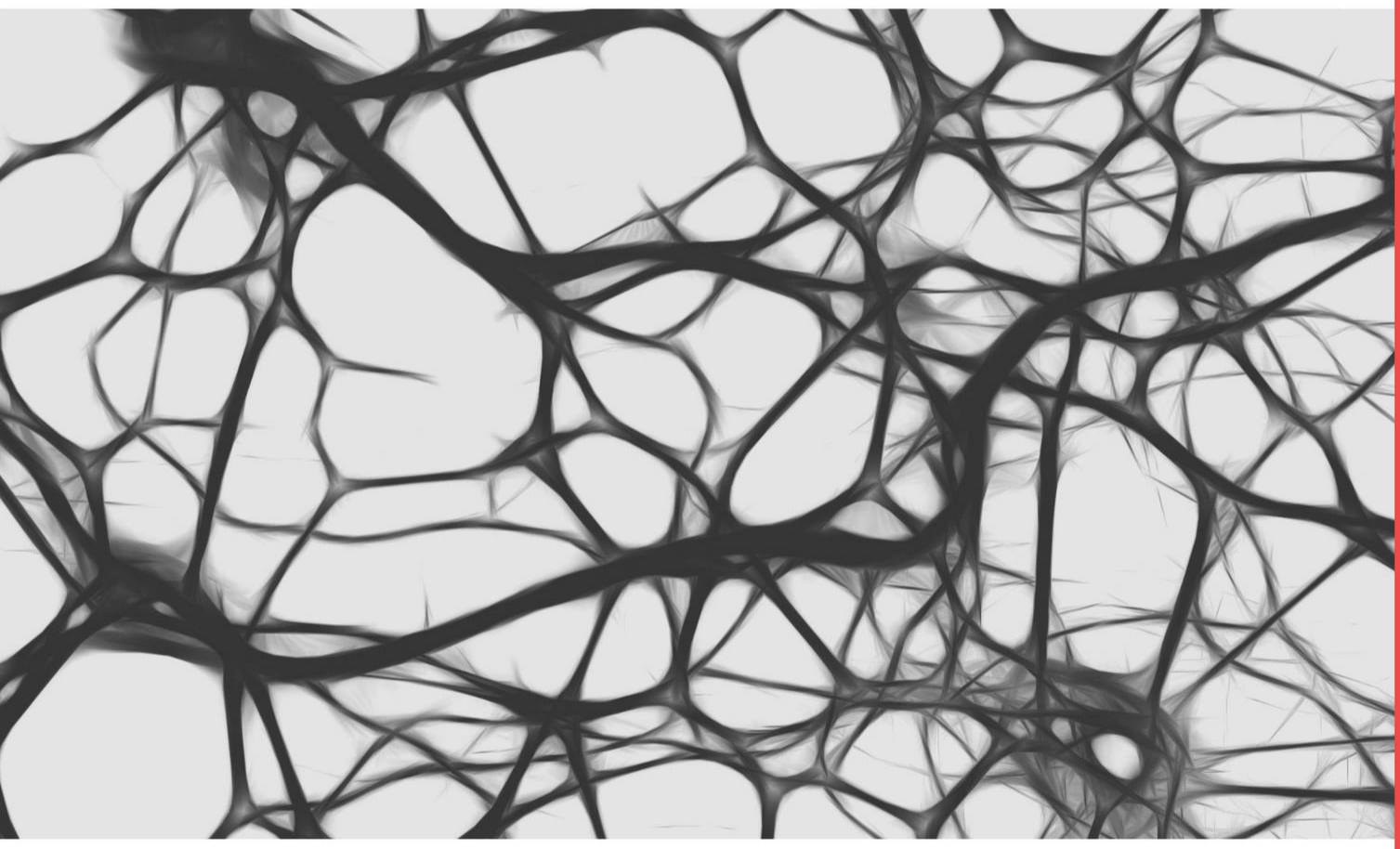
POPULATION RANGE EXPANSIONS, WITH MATHEMATICAL APPLICATIONS TO INTERACTING SYSTEMS AND ANCIENT HUMAN GENETICS

Víctor López de Rioja

UNIVERSITAT DE GIRONA

Universitat de Girona

DOCTORAL THESIS



Population range expansions, with mathematical applications to interacting systems and ancient human genetics

Víctor López de Rioja

2018

DOCTORAL THESIS

Population range expansions, with mathematical applications to interacting systems and ancient human genetics

Víctor López de Rioja

2018

DOCTORAL PROGRAMME IN THE ENVIRONMENT

Supervised by:

Joaquim Fort Viader

Neus Isern Sardó

Thesis submitted in fulfillment of the requirements to obtain the degree of Doctor of the University of Girona

Publications derived from this thesis

Three original papers have derived from this thesis which have been published in peer-reviewed journals with impact factors within the first quartile, according to the Journal Citation Reports (JCR). Chapters 3-5 are exact transcriptions of the contents of these publications, which are also included as Appendix B.

The complete references of the three papers comprised in this thesis and the impact factors of their journals are:

de Rioja VL, Fort J, Isern N. Front propagation speeds of T7 virus mutants. *Journal of Theoretical Biology* **385** (2015) 112–118. DOI: 10.1016/j.jtbi.2015.08.005. Impact factor 2.049. Journal 14 of 56, quartile 1, category *Mathematical and Computational Biology* (JCR 2015).

de Rioja VL, Isern N, Fort J. A mathematical approach to virus therapy of glioblastomas. *Biology Direct* **11** (2016) 1-12. DOI: 10.1186/s13062-015-0100-7. Impact factor 2.856. Journal 18 of 84, quartile 1, category *Biology* (JCR 2016).

Isern N, Fort J, de Rioja VL. The ancient cline of haplogroup K implies that the Neolithic transition in Europe was mainly demic, *Scientific Reports*, **7** (2017) 11229, 1-10. DOI:10.1038/s41598-017-11629-8. Impact factor 4.259. Journal 10 of 64, quartile 1, category *Multidisciplinary Sciences* (JCR 2016).

Acknowledgements

Undertaking this PhD has been a real challenge and a very enriching experience for me and it would not have been possible without the support and guidance that many people have given to me during these last three years.

First, I would like to thank my supervisors Dr. Joaquim Fort and Dr. Neus Isern for all the support and confidence they gave to me. Without their instruction and constant feedback this PhD would not have been feasible.

I gratefully acknowledge the funding received towards my PhD from Ministerio de Economía, Industria y Competitividad (Grants SimulPast-CSC-2010-00034, FIS-2009-13050, FIS-2012-31307) which supported my first months of research and also the different stays in congresses and participations in workshops in which I took part in during these years of scientific research.

This thesis has been also possible thanks to the Universitat de Girona fellowship program (Beca de Recerca UdG) which financed most of the time I have dedicated to it.

I would like to thank all the professors and colleagues from the Physics Department in Universitat de Girona too, and especially Aarón and Alex with whom I shared workplace in university (and who obtained their PhD degree during these years, congrats!).

Finally, I owe my greatest gratitude to my family, specially my mother, father and sister. Also, I would like to mention my uncle Joan and my grandma, who passed away some time before I could finish this thesis. And of course, thanks to all my friends and life partner with whom I shared so many moments: at the Physics' faculty, in sociocultural projects as Fills de la Flama or Can Batlló, in different musical projects as Rage to Antonio Machine or Föam and a very large etcetera. In some way, all these people have made me accomplish the redaction of these pages, especially Adri, Alaia, Artur, Boga, Carlo, Clara, Costa, Ger, Jordi, Laura, Ligia, Maneu, Marcel, Miki, Oscar, Pau, René, Rocío, Tarra, Torru and Xavi.

Contents

List of Figures

List of Tables

Abbreviations

aDNA: ancient DNA

B: bacterial cells

BAC: Baalberge culture

BCE: before the Common Era

BEC: Bernburg culture

BHK: baby hamster kidney

BP: before present (years BP are computed as years before year 1950 AD)

cal: calibrated

dpi: days post-infection

E. coli: *Escherichia coli*

F: farmer

GBM: glioblastoma

GFP: green fluorescent protein

HG: hunter-gatherer

hpi: hours post-infection

HRD: hyperbolic reaction-diffusion

KCN: potassium cyanide

LBK: Linearbandkeramik (German) or Linear Pottery culture (English)

MOI: multiplicity of infection

mtDNA: mitochondrial DNA

N (in context of human ancient genetics): farmer with haplogroup K

PC: Principal Component

PWC: Pitted Ware culture

PDE: partial differential equation

PFU: plaque-forming unit

PPNB: Pre-Pottery Neolithic B

PRD: parabolic reaction-diffusion

SCG: Schöningen group

SMC: Salzmünde culture

TRB: Trichterbecherkultur (German) or Funnelbeaker culture (English)

VSV: Vesicular Stomatitis Virus

Y-DNA: DNA from the Y chromosome

X (in context of human ancient genetics): farmer without haplogroup K

Abstract

Complex systems are composed mainly of various interconnected parts or dynamics. Because of this, the parts interact and the system develops new properties that cannot be explained by considering a single isolated part. This thesis focuses on a specific field of the whole possible range that encompasses complex systems, namely population dynamics. By resorting to reaction-diffusion-interaction equations, we can explain in an analytical and computational way the spatio-temporal changes of the different populations that interact between them. In this thesis, a single mathematical basis is used for the following three applications, which at first glance seem very different.

The first model presented in this thesis (Chapter 3) studies the dynamics of viral infections. Due to the numerous experimental data available, we have studied various bacteriophage T7 mutants infecting *E. coli* host bacteria. By incorporating the delay time in the terms of diffusion and reaction, as well as new mathematical terms that describe the behavior of infected cells in a biologically correct way, we obtain a model that agrees better than previous ones with the observed front propagation speeds in several strains of the T7 virus.

Chapter 4 proposes several mathematical models on the dynamics of oncolytic viruses. Oncolytic viruses are those that infect mainly tumor cells. They are applied in some experimental medical treatments of cancer. When a virus infects a tumor cell, it reproduces within it. Finally, a great number of viruses are released when the infected cell dies (lysis). Our models are focused on the spread of Vesicular Stomatitis Virus (VSV) in glioblastomas, the most aggressive brain tumors. Improvements are incorporated into the equations of a model already proposed, and comparison is performed with the results observed *in vitro*. The only model capable of efficiently explaining the dynamics of the system takes into account the delay time for the diffusion and also for the reaction processes.

Finally, Chapter 5 explains and compares quantitatively, for the first time, Neolithic DNA samples with mathematical simulations based on reaction-diffusion methods. It is believed that farming and stockbreeding (the Neolithic) spread across Europe with a progressive migration of the first farmers from the Near East, leading to the spread of new haplogroups (variations of the human DNA) that were absent in European indigenous populations (hunter-gatherers). We have performed a detailed bibliographical search to gather a database with all 513 Neolithic individuals in Europe, Anatolia (Turkey) and Syria older than 3,000 years before the Common Era (BCE), such that their mitochondrial DNA (mtDNA) has been published. With these data, it is possible to calculate what percentage of the Neolithic population had each haplogroup, in different places and times. Focusing on haplogroup K, which displays a decrease with increasing distance to the origin of the Neolithic expansion, Chapter 5 builds a computational model that takes into account the two mechanisms of Neolithic diffusion, namely demic and cultural. The simulations show that to correctly explain the genetic cline, the transition would have been basically due to population movement (demic diffusion) and only 2% of Neolithic farmers would have interbred with hunter-gatherers or taught new techniques to them (cultural diffusion).

Resum

Els sistemes complexos es formen principalment de diverses parts o dinàmiques connectades entre elles o entrellaçades. A causa d'aquesta interconnexió, les parts interaccionen i el sistema pot desenvolupar noves propietats que no podrien ser explicades a partir d'una sola part aïllada. Aquesta tesi s'enfoca en un camp específic de tot el possible ventall que engloben els sistemes complexos: la dinàmica de poblacions. Gràcies a les equacions de reacció-difusió-interacció, podem explicar de manera analítica i computacional l'evolució espaciotemporal de diferents poblacions que interactuen entre elles. En aquesta tesi, una mateixa base matemàtica s'utilitza per a les tres següents aplicacions que, aparentment, són molt diferents.

El primer model que presenta aquesta tesi (Capítol 3) estudia la dinàmica de les infeccions víriques. Degut a l'elevat nombre de dades experimentals disponibles, hem estudiat el bacteriòfag T7 infectant el bacteri *E. coli*. Gràcies a la incorporació del temps de retard en els termes de difusió i reacció, així com de nous termes matemàtics que descriuen de manera biològicament correcta la dinàmica de les cèl·lules infectades, hem aconseguit un model que mostra un millor acord que els anteriors amb les velocitats de propagació observades en diferents cepes del virus T7.

El Capítol 4 està dedicat a diferents models matemàtics de dinàmica de virus oncolítics. Els virus oncolítics són aquells que infecten principalment cèl·lules tumorals. Es fan servir en alguns tractaments mèdics experimentals de càncer. Després que un virus infecta una cèl·lula tumoral es reprodueix dins seu. Finalment s'allibera un gran nombre de virus quan la cèl·lula infectada mor (lisi). En concret, els models s'apliquen a l'expansió del Vesicular Stomatitis Virus (VSV) en glioblastomes, els tumors cerebrals més agressius. S'incorporen millores en les equacions d'un model ja proposat, i es compara amb els resultats observats *in vitro*. Es troba que l'únic model capaç d'explicar de manera eficient la dinàmica del sistema té en compte el temps de retard per als processos de difusió i també de reacció.

Finalment, el Capítol 5 explica i compara per primera vegada, d'una manera quantitativa, les mostres d'ADN neolític amb simulacions matemàtiques basades també en els mètodes de reacció-difusió. Es creu que l'agricultura i la ramaderia (el neolític) es van propagar per Europa amb una migració progressiva dels primers agricultors des d'Orient Proper i que, amb ells, es van propagar nous haplogrups (variacions trobades en l'ADN humà) que no existien en les poblacions europees autòctones (caçadors-recol·lectors). Hem portat a terme una cerca bibliogràfica minuciosa per recopilar una base de dades amb tots els 513 individus neolítics a Europa, Anatòlia (Turquia) i Síria anteriors a l'any 3.000 a.C. (abans de Crist), tals que el seu ADN mitocondrial (mtDNA) ha estat publicat. Amb aquestes dades, es pot calcular quin percentatge de la població neolítica tenia cada haplogrup, en diferents llocs i èpoques. Centrant-se en l'haplogrup K, el qual mostra una disminució respecte a la distància a l'origen de l'expansió neolítica, al Capítol 5 es construeix un model computacional que té en compte els dos mecanismes de difusió neolítica: dèmica i cultural. Les simulacions mostren que per a poder explicar correctament la clina genètica, la transició hauria d'haver estat bàsicament deguda al moviment de població (difusió dèmica) i només el 2% dels agricultors neolítics s'haurien aparellat amb caçadors-recol·lectors o els haurien ensenyat les tècniques agrícoles (difusió cultural).

Resumen

Los sistemas complejos se componen principalmente de diversas partes o dinámicas conectadas entre sí o entrelazadas. Debido a esta interconexión, las partes interaccionan y el sistema puede desarrollar nuevas propiedades que no podrían ser explicadas a partir de una sola parte aislada. Esta tesis se enfoca en un campo específico de todo el posible abanico que engloban los sistemas complejos: la dinámica de poblaciones. Gracias a ecuaciones de reacción-difusión-interacción, podemos explicar de manera analítica y computacional el cambio espacio-temporal de distintas poblaciones que interactúan entre ellas. En esta tesis, una misma base matemática es utilizada para las tres siguientes aplicaciones que, a primera vista, parecen muy distintas.

El primer modelo que presenta esta tesis (Capítulo 3) estudia la dinámica de las infecciones víricas. Debido al alto número de datos experimentales disponibles, estudiamos el bacteriófago T7 infectando a la bacteria *E. coli*. Gracias a incorporar el tiempo de retraso en los términos de difusión y reacción, así como nuevos términos matemáticos que describen de manera biológicamente correcta la dinámica de las células infectadas, conseguimos un modelo que muestra un mejor acuerdo que los anteriores con las velocidades de propagación observadas en diferentes cepas del virus T7.

El Capítulo 4 propone diferentes modelos matemáticos sobre la dinámica de virus oncolíticos. Los virus oncolíticos son aquellos que infectan principalmente células tumorales. Se aplican en algunos tratamientos médicos experimentales de cáncer. Cuando un virus infecta a una célula tumoral, se reproduce en su interior. Finalmente un gran número de virus salen cuando la célula infectada muere (lisis). En concreto, los modelos se aplican a la expansión del virus Vesicular Stomatitis Virus (VSV) en glioblastomas, los tumores cerebrales más agresivos. Se incorporan mejoras en las ecuaciones de un modelo ya propuesto, y se compara con los resultados observados *in vitro*. Se encuentra que el único modelo capaz de explicar de manera eficiente la dinámica del sistema tiene en cuenta el tiempo de retraso para los procesos de difusión y también de reacción.

Por último, el Capítulo 5 explica y compara por primera vez, de una manera cuantitativa, las muestras de ADN neolítico con simulaciones matemáticas basadas también en los métodos de reacción-difusión. Se cree que la agricultura y ganadería (el neolítico) se propagaron por Europa con una migración progresiva de los primeros agricultores desde Oriente Próximo y que, con ellos, se propagaron nuevos haplogrupos (variaciones encontradas en el ADN humano) que no existían en las poblaciones europeas autóctonas (cazadores-recolectores). Hemos llevado a cabo una búsqueda bibliográfica minuciosa para recopilar una base de datos con todos los 513 individuos neolíticos de Europa, Anatolia (Turquía) y Síria anteriores al año 3.000 a.C. (antes de Cristo), tales que su ADN mitocondrial (mtDNA) ha sido publicado. Con estos datos, se puede calcular qué porcentaje de la población neolítica tenía cada haplogrupo, en diferentes lugares y épocas. Centrándose en el haplogrupo K, el cual muestra una disminución respecto a la distancia al origen de la expansión neolítica, en el Capítulo 5 se construye un modelo computacional que tiene en cuenta los dos mecanismos de difusión neolítica: démica y cultural. Las simulaciones muestran que para poder explicar correctamente la clina genética, la transición debería haber sido debida básicamente al movimiento de población (difusión démica) y tan solo el 2% de los agricultores neolíticos se habrían apareado con cazadores-recolectores o les habrían enseñado nuevas técnicas (difusión cultural).

PART I

Introduction, objectives and methods

1. Introduction

This thesis focuses on the quantitative analysis of biophysical systems where dispersion and reaction processes coexist (both in microbiological and in human populations). The theories usually applied to describe such systems are called reaction-diffusion theories, and their basic formulation will be reviewed in Sec. 1.2 below. Reaction-diffusion theories are commonly used in chemistry to describe systems where one or more substances may chemically react and diffuse in the medium containing them, leading to changes of their densities in space and time. A well-known example is the propagation of combustion flames [1]. Reaction-diffusion processes also take place in physical systems, e.g. superconductors [2], as well as in many biological and social phenomena that are of importance in archaeology [3], virology [4], genetics [5] and linguistics [6]. In biological systems, the diffusion process describes random migratory movements of the individuals, and the reaction process can include both population growth (births and deaths) as well as interactions between several populations or species (e.g., predator-prey, competition, or symbiotic interactions).

Each of the three main Chapters of this thesis (Chapters 3-5) reproduces a research paper devoted to describe mathematically a specific biological system [7, 8, 9]. All three systems share a common feature, namely that various species or populations change their number densities by reproducing, interacting with other species and spreading throughout the medium where they live. These three papers [7, 8, 9] try to contribute and improve previous work on modelling this type of reaction-diffusion systems in three separate fields of study, namely the spread of viruses in a bacterial medium (Chapter 3) [7], the spatiotemporal dynamics of oncolytic viruses injected to defeat cancerous tumors (Chapter 4) [8], and the study of the variation in space and time of the percentage of a Neolithic genetic marker in ancient human populations (Chapter 5) [9].

This introduction contains the background on each of these three systems (Sec. 1.1), an overview of population dynamics models previously applied to such problems (Sec. 1.2), and a summary of the models applied in this thesis (Sec. 1.3).

1.1. Three applications of reaction-diffusion processes

This section provides a brief introduction to the three systems studied in this thesis. It also includes discussions on their scientific interest, as well as on the features that are relevant for this thesis.

1.1.1. Virus infection fronts

Virus growth dynamics differs significantly from that of cellular organisms. The so-called one-step growth experiments were devised by physicist Max Delbrück. In these experiments, viruses are distributed *homogeneously* in a medium of susceptible or host cells (i.e., cells that can be infected and killed by the viruses). The viruses first adsorb to the cells and, some time later, it is observed that all the viruses reproduce almost at once (i.e., at 'one step'), so the population number does not increase exponentially as for cellular organisms [10]. For this and related work, Delbrück was awarded the Nobel prize in Medicine or Physiology in 1969.

An important class of *non-homogeneous* systems is obtained by injecting viruses in a small region of a medium containing host cells, i.e. cells that can be infected by the viruses. Then it is observed that a circular region of dead cells (called a plaque) gradually grows in size [11]. For many virus-host systems, viruses can only infect hosts that are reproducing, thus the plaque growth stops spontaneously when host reproduction ceases due to exhaustion of nutrients [11] (unless fresh host is added, but this complicates accurate experimental measurements [12]). On the other hand, the plaques of some virus-host systems (e.g., those of virus T7 infecting *E. Coli* cells) grow without any bound (other than the dimensions of the experimental setup) because cell infection takes place even without host reproduction [12]. In such systems, it is easy to measure the plaque radius as a function of time, even after the plaque becomes visible to the naked eye, and this makes it possible to perform accurate estimations of the speed of plaque growth [12].

In Chapter 3 we shall consider bacteriophages (i.e., viruses that infect bacterial cells). The experimental technique consists of distributing many susceptible bacteria homogeneously on a disk-shaped surface (Petri plate) containing nutrient agar, which immobilizes the bacteria, so they do not diffuse [12]. A few viruses are then injected at the center of the plate and the so-called lytic cycle begins, i.e. each virus spreads through nearby bacteria, infects one of them, reproduces inside it until this cell dies (lysis) and the virus progeny leave it. This cycle repeats many times, leading to a growing region of killed (or lysed) bacteria, i.e. a plaque (see Fig. 1.1 for a scheme). Between the adsorption of a virus into a cell and the exit of the virus progeny from it, there is an elapsed time which can range from minutes to hours. During this time interval, the virus replicates within the host cell. Thus, during this time interval neither the original virus nor its progeny moves. This time interval is very important in this thesis. It is called the delay time, latent period, eclipse time or lag time, and is usually denoted by τ.

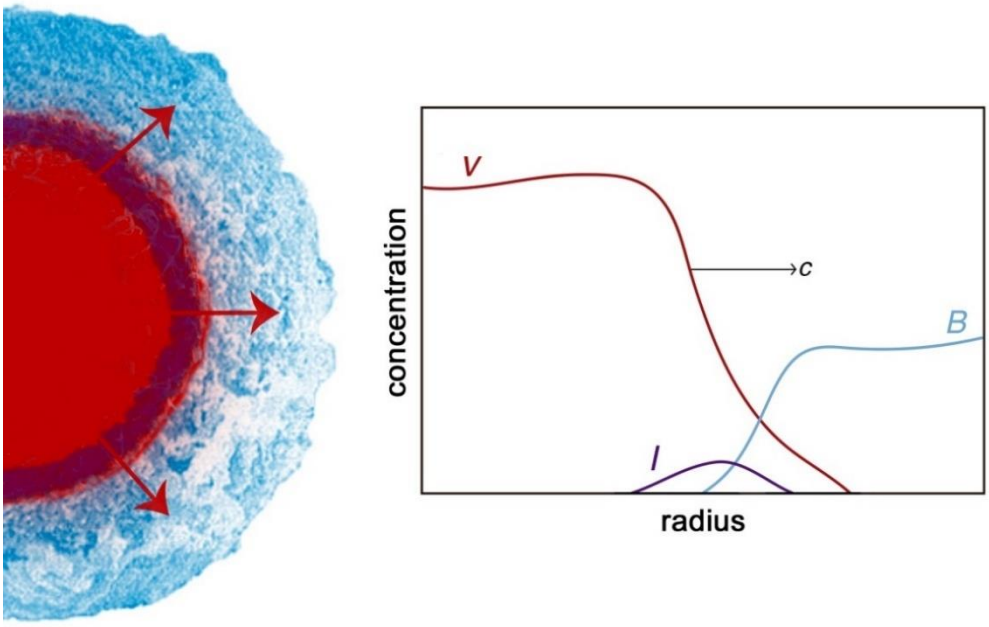


Figure 1.1 Left: Diagram representing the virus (red) propagation front through the host cells (blue). Right: Dimensionless radial profiles of the concentrations of viruses $\boldsymbol{V}$ (red), infected cells $\boldsymbol{I}$ (purple) and uninfected cells $\boldsymbol{B}$ (blue) in a growing plaque. All three profiles move with the same speed *c*. Such a profile for $\boldsymbol{V}$ is called a front (or sometimes a wavefront). The profile for $\boldsymbol{I}$ is called a pulse. Some authors also call them travelling waves.

The edge of a plaque has a well-characterized and constant speed (Fig. 1.1). The speed of the front can be measured experimentally [12]. It can be also calculated from reaction-diffusion sets of equations (see Sec. 1.2.3 below). Such equations consider the three processes at work, namely (i) diffusion of viruses, (ii) adsorption or entrance of viruses into host cells (which will become infected), and (iii) reproduction of viruses (the new offspring being released after the host cell has been infected for a certain time τ). These processes are depicted in Fig. 1.2.

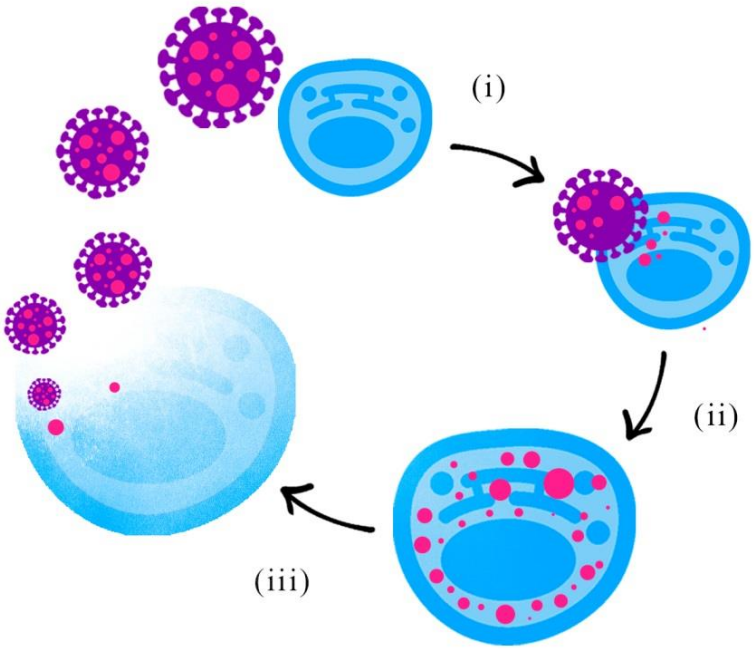


Figure 1.2 Schematic of a virus lytic cycle (viral replication): (i) search, attachment and entrance to the susceptible cell, (ii) production of new viruses while inside the cell and (iii) burst and release of the new viruses.

Koch developed a first model in 1964 which, based on heuristic arguments, assumed that the rate of growth of the plaque should be constant and dependent only on the viral diffusivity and the lag time of viral replication [13] (the model due to Koch and its limitations will be described at the end of Sec. 1.2.1.). Later other authors have developed more accurate and realistic models, with rigorous derivations that take into account the rates of virus adsorption and cell lysis. We can highlight the work by Yin and co-workers [12, 14, 15, 16, 17, 18], who performed many experiments and incorporated new kinetic parameters into their renewed reaction-diffusion equations. They computed the speed of travelling-wave solutions (i.e., the speed of plaque growth) analytically [14] and also performed numerical simulations [17]. They studied the dynamics and evolution of several strains of the bacteriophage T7 infecting *Escherichia coli* (*E. coli*). We stress that they could perform accurate experimental measurements due to the fact (mentioned above) that, unlike the plaques of many viruses that stop growing (when cells reach a stationary state, i.e. when there is no net cell reproduction), plaques of T7 on a plate continue to grow in size indefinitely (until the edge of the plate is reached) [12]. A decade later, it was noted that the T7-*E. coli* speed resulting from the models due to Yin and co-workers was much faster than the experimental one [19]. For this reason, and Fort and Méndez [4] took into account the latent period τ during which viruses do not move, by using a time-delayed equation. The computed wave speeds then agreed perfectly with the observed ones [4]. However, the term including the latent time is not biologically intuitive. Chapter 3 explains this

problem and tries to fix it, by means of a new term that we consider mathematically and biologically more appropriate. In the paper reproduced in Chapter 3 [7], we consider only T7 viruses infecting *E. coli* for definiteness. However, this is not the only system to which our results can be applied. Indeed, in the next section and in the paper in Chapter 4 [8] we consider Vesicular Stomatitis Viruses (VSVs). These viruses infect mammalian or insect cells (not bacteria) [18, 20, 21]. For VSV infections, the cells do not burst. But again, there is a delay time between the adsorption of a virus by a cell and the release of the virus progeny. For this reason, our models are also relevant to VSV infections.

1.1.2. **Oncolytic virotherapy**

Cancer represents a threat of utmost importance for public health worldwide. Traditional treatments are surgery, radiotherapy and chemotherapy. But despite important advances, certain forms of cancer do not respond well to these traditional treatments. For example, glioblastomas (GBM) are highly malignant brain tumors (median survival <15 months [22]) and their treatment will require innovative approaches. One of them is oncolytic virotherapy [23]. In this technique, viruses that infect cancerous cells are injected into a tumor [24, 25]. Some strains of Vesicular Stomatitis Viruses (VSVs) are candidates for such therapies due to their effectiveness infecting cancer cells, without damaging healthy cells nearby [26, 27]. Some strains of VSVs replicate in tumoral cells and kill them, by following a cycle like that represented in Fig. 1.2. The hope is, then, that VSVs could be able to infect all the cancerous cells and defeat the tumor.

The original idea of treating tumors with viruses arose already in the 20th century [28], but research was then limited because of the difficulty to obtain sufficiently effective viruses to infect tumors, without killing healthy cells. Nowadays, due to the rapid development of genetic engineering, scientists can create viruses with improved skills, e.g. with increased tumor-cell selectivity, more effective replication inside infected cells, and enhanced oncolytic activity leaving the normal cells unharmed. For these reasons, interest in cancer treatments with viruses has reappeared tremendously during the last decade [29, 30, 31].

In parallel, mathematical models of cancer treatment with viruses have been developed in the last years. This biophysical system is very similar to that described in the previous section, in the sense that a virus is injected into a tumor, and spreads through the population of cancer cells (for *in vitro* experiments, the spread takes place in a basically two-dimensional geometry). Because of the similarities with the system in the previous section, the virus-tumor travelling wave can also be described using reaction-diffusion equations. However, in a virus-tumor system, the tumor is also expanding. For this reason, contrarily to the representation in Fig. 1.1, there is an additional travelling wave of tumor cells (i.e., an expansion of the outer circle in Fig. 1.1). It is therefore necessary to model tumor growth (i.e., the spread of the blue region in Fig. 1.1). In contrast to solid tumors [32, 33], the cells of glioblastomas diffuse. Therefore, the growth of glioblastomas can be described using reaction-diffusion equations (see, e.g., Ref. [34] and citations therein). More information on glioblastoma growth is given in Sec. 2.1.6 below.

Wodarz and co-workers have applied previous, more general work on virus-cell systems (as reviewed by Nowak and May [35]) to the more specific case of virus-tumor systems [36, 37, 38, 39, 40]. Those works considered only non-spatial systems, i.e. without diffusion (so they are based on ordinary differential equations). Spatial models (based on partial differential equations) have been also

considered [41, 42, 43, 44, 45, 46]. They are reviewed in Sec. 1.2. In Chapter 4 we use spatial models to describe the propagation of the virus infection front (red region and arrows in Fig. 1.1) as well as that of the tumor cell front (blue region in Fig. 1.1) [8].

1.1.3. DNA clines and the nature of the Neolithic spread

The Neolithic is defined as an economic system based on farming and stockbreeding, as opposed to the hunting and gathering economy practiced during the Paleolithic (including its final phase, which is often called the Mesolithic). The adoption of this new economic system was a major transition for humankind. The earliest Neolithic sites appeared about 11,000 yr BCE in the Near East, i.e. Israel, Jordan, Iraq, Syria, etc.). From there, the Neolithic spread to Anatolia (present-day Turkey), next to Greece, and then along two main routes: northwards to Central Europe and westwards to the Iberian Peninsula (along the Mediterranean coast). It arrived at around 6,000 yr BCE to southeastern Europe [47], at about 5,500 yr BCE to Portugal [48], and at around 4,000 yr BCE to the British Islands and Scandinavia [47]. This process radically changed the environment and led to an increase in population densities, as well as to new forms of social organization [49]. An important question is whether the spread of the Neolithic was due to demic diffusion (dispersal of farming populations), to cultural diffusion (incorporation of hunter-gatherers into the farming populations, either via interbreeding with farmers and/or becoming farmers by acculturation), or to a combination of both mechanisms. These processes would have left different footprints on the genetics of the first farmers. For this reason, the study of the Neolithic gene pool may be crucial to achieve a better understanding of the process. We deal with this problem in Chapter 5 [9].

In 1971, archaeologist A.J. Ammerman and geneticist L.L. Cavalli-Sforza analyzed the dates of the early Neolithic European sites that had been discovered and dated at the time. In this way, they were the first to provide a statistically sound estimation of the speed of the spread of the Neolithic in Europe [50]. Their result for the observed speed, $c_{obs} = (1.0 \pm 0.2)$ km/year [50, 51], should be understood as the average spread rate from a presumed origin in Jericho (the Neolithic site in the Near East which yielded the highest correlation coefficient, namely $r = 0.89$). Ammerman and Cavalli-Sforza also noted regional variations (a slowdown in the Alps and a faster spread along the Mediterranean) [50] and later used spatial interpolation techniques to generate isochrone maps of the spread of farming in Europe [52]. By comparing the observed rate mentioned above, i.e. (1.0 ± 0.2) km/year [50, 51], to a similar value resulting from a reaction-diffusion model due to Fisher (and explained in Sec. 1.2.1 below), they also postulated that the propagation of the Neolithic could have been due mainly to the dispersal of farming populations (demic diffusion) rather than acculturation or interbreeding (cultural diffusion) [50, 53]. And, crucially, they suggested that the spread pattern observed from archaeological data (i.e., younger Neolithic sites with increasing distance from the Near East) might be detected in the genetics of modern human populations, in the form of spatial variations (clines) in the frequencies of some alleles (i.e., variations of a gene) with a maximum (or a minimum) in the Near East (due to interbreeding with hunter-gatherers) [50, 53]. Note that, whereas Ammerman and Cavalli-Sforza proposed that demic diffusion could have been the most important mechanism spreading the Neolithic, they also suggested a clear role for cultural diffusion, because in their proposal interbreeding between farmers and hunter-gatherers would have led to the formation of genetic clines, although they had not been yet observed by then. Some years later, their prediction was impressively confirmed by synthetic genetic maps of modern Europeans [54, 55]. We mention

that other massive migrations (such as the arrival of modern humans about 50,000 yr BCE) have surely had effects on such modern genetic maps (see below), and that other possible mechanisms leading to genetic clines have been proposed (see, e.g., Ref. [56]). However, the Neolithic transition is widely considered as a relevant cause of the observed modern genetic pattern [57, 58].

The Y chromosome is one of the two sex chromosomes (X and Y) found in the cellular nucleus, whereas mitochondrial DNA (mtDNA) is located in cellular mitochondria (organelles located outside the nucleus). Many studies on human population genetics have analyzed the Y chromosome and mtDNA [59]. These two types of DNA are uniparentally inherited (the Y chromosome is present only in men and is inherited from the father, whereas mtDNA is present both in men and women and is inherited from the mother) [59, 60, 61]. Different migratory behaviors in women and men will lead to different mtDNA and Y-chromosome patterns [62]. Although most of the DNA in the Y chromosome (Y-DNA) and the mtDNA is transmitted without changes to the next generation, the offspring will often inherit the DNA information with some modifications. The individual sequences of Y-DNA and mtDNA are known as haplotypes, but similar haplotypes with a common ancestor are usually grouped under haplogroups (which are therefore groups of sequences with a common ancestor), which are useful for the study of genetic composition of the early Neolithic populations of the genetic effect of the Neolithic spread.

Since four decades ago, studies on genetic markers of *present* populations have tried to shed some light into the genetic effects of the Neolithic transition [54, 63]. Fifteen years ago, Semino et al. [60] studied the Y chromosomes of 1,007 modern Europeans. Their analysis suggested that about 80% of the European human gene pool dates back to Paleolithic, and only 20% of European modern DNA can be included into the Neolithic package, in agreement with previous mtDNA studies [64, 65] and work on so-called classical markers (blood groups, Rh, HLA-B, etc.). Indeed, classical markers suggested that the spatial function that accounts for most of the spatial variation of present European genetic frequencies (the so-called first principal component (PC)) corresponds to the Neolithic transition, and accounts for about 28% of the total variation [54, 55, 66]. It is also very interesting that Semino et al. [60] identified as Neolithic markers those with higher correlations with the first PC and found that their frequencies are higher for Mediterranean populations than for non-Mediterranean ones. They interpreted this as the result of stronger demic diffusion along the Mediterranean than in northern Europe, a possibility that has been suggested independently by archaeologists [67, 68, 69, 70]. However, the genetics of modern Europeans have been surely affected by subsequent substantial population movements (besides Paleolithic and Neolithic range expansions). For example, the well-known Bronze-age migrations from the Urals that possibly spread the Indo-European languages had strong genetic effects [71]. Therefore, it is clear that in order to disentangle quantitatively the effects of demic and cultural diffusion in the spread of the Neolithic, it is better to use ancient (rather than modern) DNA. However, until recent years it has been very difficult to determine the DNA of ancient individuals, due to the degradation of DNA molecules.

In year 2012, a study proposed that demic diffusion had a more important effect than cultural diffusion on the spread rate of the Neolithic [3]. In contrast, the genetic consensus at the time was that cultural diffusion had been more important than demic diffusion [60, 72, 73, 74]. However, that proposal [3] was based on purely on archaeological, non-genetic data, and there is no general theory showing that the primacy of demic over cultural diffusion concerning the spread rate necessarily implies its primacy concerning the genetic pool.

As mentioned three paragraphs above, a mtDNA haplogroup is a group of several sequences (or collections of allelels, which are called haplotypes) of mtDNA with a common ancestor. Each individual has a single mtDNA haplogroup. Haplogroups usually date back thousands of years, when a mutation occured (thus leading to the first individual with the haplogroup considered). Subsequent mutations lead to new haplogroups (which are subgroups, or subclades, of the older haplogroup), e.g. A1a and A1b are subclades of haplogroup A1.

The possibility to analyze *ancient* DNA has become a reality only during the last 15 years, due originally to technological advances in DNA sequencing and, more recently, to the identification of some parts of the human body (e.g., the petrous bone in the inner ear) giving excellent DNA yields [75]. The first relevant study using *ancient* DNA was published in year 2005. It detected the N1a haplogroup in a surprisingly high proportion (6 of 24 skeletons) of early farmers of the LBK culture in Germany, Austria and Hungary [76]. N1a is very rare in Europe at present, so Ref. [76] suggested that present Europeans descend from local hunter-gatherers rather than incoming farmers, i.e. that the Neolithic spread in Europe had been mainly cultural. However, this conclusion has been ruled out by the results of later ancient DNA studies [77], which indicate that the now common mtDNA haplogroups H, T, K and J are absent in HGs. The genetic pool of European hunter-gatherers (HG) comprises exclusively U mtDNA haplogroup lineages (U, U4, U5 and U8), which are rare at present in Europe [77]. Early farmers from Europe, Syria and Anatolia are characterized by various mtDNA haplogroups including N1a, T2, K, J, HV, V, W, and X (also known as the 'Neolithic genetic package' [78]).

Since year 2010 [79], the study of ancient DNA has gone a step further by performing genome-wide studies. This makes it possible to determine not only the haplogroup of, e.g., the mtDNA, but also the presence or absence of millions of mutations in the individual considered [79]. Genome-wide studies have allowed further inferences on the demic or cultural nature of the Neolithic expansion. For example, Mathieson et al. [80] performed a genome-wide study of the largest ancient DNA (aDNA) dataset assembled by the time. That database had 230 individuals. Of these, 28 were Anatolian Neolithic farmers, and their genome-wide DNA was reported in Ref. [80] for the first time. The results clearly support a demic Neolithic expansion, with very little cultural diffusion. Indeed, Mathieson et al. [80] estimated that early Neolithic European farmers had a genetic Anatolian component larger than 90%, and the rest (below 10%) was identified as hunter-gatherer ancestry. This implies that the *modern* DNA work summarized above [64, 65, 60] had erroneously identified the non-Neolithic component as a Paleolithic one, whereas *ancient* DNA [80, 81, 82] indicates that it must be mainly due to post-Neolithic migrations [83, 74, 84, 85]. It also implies that demic diffusion was very important in the spread of the Neolithic, as originally proposed by Ammerman and Cavalli-Sforza [50, 53]. Very recently, ancient genetics has addressed another very important question, namely whether the first farmers that brought agriculture to different regions of Europe were all derived from a single source population or from several ones. Genome-wide results indicate that early Neolithic farmers from Iberia (Epicardial culture), central Europe (LBK culture), the Balkans and Anatolia [86], as well as those from Britain [87], are all closely related. This provides strong support for a single migration from Anatolia.

For the purposes of this thesis, it is important to stress the following point. In order to determine the percentage of farmers involved in cultural transmission, we think that the best approach is to consider a single marker that has not been apparently affected by other processes (such as selection,

mutation, drift, etc.). In this way, we do not have to introduce any additional unknown parameter values related to those other processes. Such a marker will be studied in Chapter 5 [9], where we shall also present a model that predicts its percentage in the population as a function of position (i.e., its genetic cline). In Chapter 5 we shall show that comparing the predicted cline to the observed one makes it possible to estimate the percentage of farmers involved in cultural transmission [9]. In contrast, genome-side studies include many markers. As mentioned above, some of them can be affected by other processes besides cultural transmission (selection, mutation, drift, etc.). Thus, each marker can have a very different cline, depending on the processes that have shaped it. For example, drift can drastically reduce or increase the frequency of a marker during the propagation of a population front [56]. For this reason, in our opinion genome-wide studies such as that by Mathieson et al. [80] cannot be used to explain the shapes of specific genetic markers, neither to determine the percentage of farmers involved in cultural diffusion. For this, it seems necessary to consider the cline of a marker that has been affected by cultural diffusion, and not by any other effects. We will study (in Chapter 5) the spatiotemporal variation on the frequency of a mitochondrial Neolithic marker, namely haplogroup K, and evaluate quantitatively the fraction of the Neolithic population involved in cultural diffusion. This will be done by comparing the observed ancient cline of haplogroup K to the results of a demic-cultural reaction-diffusion model. Previous authors had performed demic-cultural reaction-diffusion studies [5, 88, 89, 90, 91, 92], although our equations are different because we use cultural transmission theory (see Sec. 1.2.5 below) [93, 94, 3, 95]. Another difference with those earlier works is that we compare (in Chapter 5) [9] to ancient rather than to modern DNA data.

1.2. Previous mathematical models

This section summarizes previous attempts to mathematically model the three systems studied in this thesis. Although all the models have been adapted to better suit the biological application they pretend to describe, it is important to emphasize that all of them are reaction-diffusion models based on an equation introduced by R. A. Fisher in the 1930s [96]. The subsections below first provide a general overview of the initial attempts to improve that original reaction-diffusion equation, and then the specific approaches to the three systems we deal with in Chapters 3-5.

1.2.1. One equation to rule them all

For a two-dimensional space (i.e., a surface), the simplest equation describing a system where diffusion and reaction processes coexist is

$$\frac{\partial p}{\partial t} = D\left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}\right) + F(p). \tag{1.1}$$

Below we give a derivation of this equation. In biological systems, $p = p(x, y, t)$ is the population density (number of individuals per unit area at position (x, y) and time time t) of the single population present in this model, D is its diffusion coefficient, and the so-called growth function $F(p)$ includes the effect of net reproduction (births and deaths). The same equation can be applied to physical and chemical systems (in the latter case, $p(x, y, t)$ is the concentration of a chemical species, and $F(p)$ includes the effect of chemical reactions). In general, the diffusion coefficient can be written as $D = \frac{\langle \Delta^2 \rangle}{2n\tau}$, where $\langle \Delta^2 \rangle$ is the mean squared displacement, τ is the time interval between two successive

jumps of a particle (or individual), and n is the number of spatial dimensions [97, pp. 78-79]. Therefore, in the special case of a two-dimensional space ($n = 2$), as in most of this thesis, the diffusion coefficient is given by $D = \frac{\langle \Delta^2 \rangle}{4\tau}$ (a proof of this equation is given in Sec. 1.2.2 below) [97, 98]. Equation (1.1) is called the Kolmogorov-Petrovsky-Piskounov (KPP) equation [99], and is classified as a parabolic reaction-diffusion (PRD) equation. This equation describes the variation in time on the population density p (left-hand side) because of a diffusion process (first term on the right-hand side), characterized the diffusion coefficient D, and a reaction process characterized by the function $F(p)$. If $F(p) = 0$, there is no reaction process and Eq. (1.1) corresponds to the so-called Fickian diffusion. This kind of diffusion is the simplest one, and assumes a random walk for individuals, in the sense that there is no preferential direction of movement [100, 101].

For the sake of completeness, let us include here a well-known derivation of Eq. (1.1) for the simplest possible case, namely one spatial dimension (i.e., $n = 1$) and a single jump distance [102, pp. 17-21] (a more complicated derivation, valid for $n = 2$ and several jump distances will be provided laer on, in Sec. 1.2.2). Consider a physical (or biological) one-dimensional system (e.g., a narrow wire or a population of birds living along a coast) such that all particles or individuals jump a distance $\pm\Delta x$ every time interval τ. Suppose that we know the number of particles at two very close points along the system at time t, namely $P(x)$ and $P(x + \Delta x)$. If the particles move randomly, after the time interval τ we can assume that half the particles initially at x will have moved across a transversal area A (located between the two points) from left to right, and half the particles initially at $x + \Delta x$ will have moved from right to left across the same unit area A. Therefore, the net flux of particles along the x axis and across the unit area A (i.e., the number of particles crossing A per unit area and unit time) can be written as

$$J_x = -\frac{1}{2}\frac{[P(x + \Delta x) - P(x)]}{A\,\tau}. \tag{1.2}$$

As mentioned above, the diffusion coefficient in one dimension is defined as $D = \frac{\Delta x^2}{2\tau}$, so we can multiply the previous equation by Δx^2 and use this equation for D. Then, since the number of particles divided by the volume ($A\,\Delta x$) is the number density of particles, $p = P/(A\,\Delta x)$, the flux Eq. (1.2) can be rewritten as

$$J_x = -D\,\frac{p(x + \Delta x) - p(x)}{\Delta x}. \tag{1.3}$$

In the limit $\Delta x \to 0$, the right-hand side is a partial derivative in space, thus we obtain

$$J_x = -D\,\frac{\partial p}{\partial x}. \tag{1.4}$$

This is called Fick's law and states that the net flux is proportional to the gradient of the concentration. The derivation of Eq. (1.1) for the special case $F(p) = 0$ follows from combining Eq. (1.4) with the law of conservation of mass (or of particle number). The latter is formalized as follows. Consider a box of volume $A\,\Delta x$ (see Fig. 1.3). Obviously, $J_x(x)A\tau$ particles will enter the box from the left and $J_x(x + \Delta x)A\tau$ will leave it to the right during the time interval. If particles are neither created nor destroyed, the number of particles per unit volume (i.e., the density of particles) in that box will increase at rate

$$\frac{[p(t+\tau)-p(t)]}{\tau} = -\frac{1}{\tau}\left[\frac{J_x(x+\Delta x)A\tau - J_x(x)A\tau}{A\,\Delta x}\right] = -\frac{[J_x(x+\Delta x)-J_x(x)]}{\Delta x}. \tag{1.5}$$

In the limit $\Delta x \to 0$ and $T \to 0$, the first and last quotients are partial derivatives, so we obtain

$$\frac{\partial p}{\partial t} = -\frac{\partial J_x}{\partial x}. \tag{1.6}$$

Using Fick's law, Eq. (1.4), into this equation, we finally obtain

$$\frac{\partial p}{\partial t} = D\frac{\partial^2 p}{\partial x^2}. \tag{1.7}$$

which is Eq. (1.1) for the special case of no reaction (i.e. $F(p) = 0$), one spatial dimension (i.e., $n = 1$) and a single jump distance (Δx). It is easy to generalize the previous derivation to more dimensions and jump distances (see, e.g., Ref. [103]). A derivation for the case most relevant to this thesis ($n = 2$ and several jump distances) is given in Sec. 1.2.2 below.

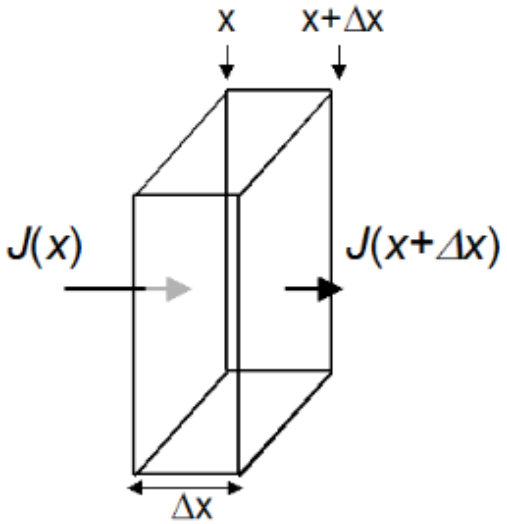


Figure 1.3 The net flux passing by the faces of a thin box of volume $\boldsymbol{A\,\Delta x}$. The surfaces normal to the $\boldsymbol{x}$ axis have area $\boldsymbol{A}$.

The last term in Eq. (1.1), i.e. $F(p)$, is simply added to take care of the population density change due to net reproduction (or to the species concentration change due to chemical reactions, in chemical systems). In 1937, the same year of publication of the KPP equation (1.1) [99], Fisher wrote a paper on the spread of advantageous genes in which he proposed a logistic function for the growth or 'reaction' term, namely [96]

$$F(p) = ap\left(1 - \frac{p}{p_{max}}\right), \tag{1.8}$$

were p_{max} is called the saturation density. Equation (1.8) is used for the following reason. Consider homogeneous systems, i.e., such that p does not depend on the spatial coordinates x, y. Then, there is no diffusion (because Eq. (1.4) yields $J_x = 0$) and Eq. (1.1) becomes simply $\frac{\partial p}{\partial t} = F(p)$. For logistic growth, this reads $\frac{\partial p}{\partial t} = ap\left(1 - \frac{p}{p_{max}}\right)$ and, if the population density is initially small ($p \approx 0$), we have $\frac{\partial p}{\partial t} \approx ap$ at early times. This implies an initial exponential growth (with rate a). However, at some point

p will become large enough so that the negative term $\left(-\frac{p}{p_{max}}\right)$ will no longer be negligible. Then the population density increases per unit time ($\frac{\partial p}{\partial t}$) will become gradually slower, until $p = p_{max}$. At this point, obviously $\frac{\partial p}{\partial t} = 0$ and the population number does not increase further. Thus, we have a self-limiting population growth, and we can interpret p_{max} as the maximum possible value of the population density is (therefore it is called the saturation density or carrying capacity). This scenario makes biological sense, because populations cannot become arbitrarily large (due to limited nutrients, space, etc.). Indeed, Eq. (1.8) is widely used in mathematical biology [104], because it is realistic for many microbiological [105] and ecological [106] systems. Using Eq. (1.8) into Eq. (1.1) yields the well-known Fisher equation [96],

$$\frac{\partial p}{\partial t} = D\left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}\right) + ap\left(1 - \frac{p}{p_{max}}\right). \tag{1.9}$$

A travelling wave (also called front or wave of advance) is, for the purposes of this subsection, a solution to a reaction-diffusion equation (e.g., Eq. (1.9)) with constant shape and speed, that describes a population invading empty space from some initial region. For Eq. (1.9), Fisher derived the following equation for the speed *c* of travelling waves

$$c = 2\sqrt{aD}, \tag{1.10}$$

which is the well-known Fisher propagation speed (a detailed derivation of Eq. (1.10) from Eq. (1.9) is included in Sec. 2.3.1 in this thesis). This was the first mathematical approach to the spreading of biological populations, and the starting framework of most subsequent modelling approaches. For the purposes of this thesis, we have found it more useful to present the results due to Fisher considering that *p* is the population density (whereas Fisher considered it to be the frequency of a mutant gene). In fact, some years later (in 1951) J. G. Skellam [107] was the first (in a paper devoted to biological invasions) to consider that *p* is the population density, although he considered that the population growth was Malthusian, namely

$$\frac{\partial p}{\partial t} = D\left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}\right) + ap. \tag{1.11}$$

This is known as the Skellam equation. Unlike that due to Fisher (Eq. (1.9)), Eq. (1.11) does not contain the last, non-linear term. It is easy to see that Eq. (1.11) corresponds to unbounded population growth, i.e. that the population density can become arbitrarily large, which is not biologically realistic. In contrast to the Fisher equation (1.9), Eq. (1.11) can be solved explicitly. For Eq. (1.9), the invading species advances at a constant speed given again by Eq. (1.10) [108].

Equations similar to Eq. (1.1) can be also applied to the spread of virus infections. As mentioned above (Sec. 1.1.1), A. L. Koch made the first attempt to describe the growth of viral plaques quantitatively [13]. Koch studied the phage T4 growing on *E. coli* and noticed that the system behaves like that of Fig. 1.2 (which describes the virus lytic cycle). Therefore, phage reproduction in a plaque requires replication of viruses inside the host cell during a time interval τ before the lysis of the host cell, leading to the subsequent then virus dispersal (with diffusion coefficient D, dependent on the virus and the medium) and adsorption to new host cells. From heuristic arguments, Koch proposed that, if the adsorption process is fast enough, the speed of plaque growth should be approximately

$c \propto \sqrt{D/\tau}$. This conclusion can be also reached using dimensional analysis, but at present it is known that, unfortunately, this approximation breaks down because in practice adsorption is not fast enough. In 1992, Yin and McCaskill developed a set of reaction-diffusion equations capable of describing in a quantitative way the spatial dynamics of a growing plaque [14]. They used three equations because they were dealing with three interacting species (viruses, healthy bacteria and infected bacteria), but these equations were based on the KPP model Eq. (1.1). Their model will be described in Sec. 1.2.3.

The Fisher equation (1.9) has important applications in archaeology and population genetics. During his postdoctoral work on bacterial genetics with R. A. Fisher in Cambridge (1949-51), L. L. Cavalli-Sforza became aware of the paper published by Fisher in 1937 [109]. Two decades later, Cavalli-Sforza was working on human genetics and became interested in the effects of prehistoric population range expansions. This led him to analyze the dates of early Neolithic sites in Europe with archaeologist A. J. Ammerman. As mentioned above (Sec. 1.1.3), in 1971 they obtained the first statistically sound estimation of the speed of the spread of the Neolithic in Europe, $c_{obs} = (1.0 \pm 0.2)$ km/year [50, 51]. Later Ammerman and Cavalli-Sforza suggested that Fisher's equation (1.9) might be adequate to describe the Neolithic spread across Europe [53]. Here the idea is that, although individual migratory movements should obey particular motivations (locations of water, fertile land, etc.), such individual preferences will be averaged if we consider many individuals and large geographical areas. Therefore, the modelling at large scales in space and time can arguably be based on reaction-diffusion equations with Fickian diffusion. Later Ammerman and Cavalli-Sforza [52] estimated the diffusion coefficient D and initial growth rate a (using data from etnhographic obervations) and calculated the farming spread rate using Fisher's speed (1.10). In fact, if using realistic values of D and a for preindustrial populations, this theoretical Neolithic front propagation speed was somehow faster than the observed value, mentioned above, namely $c_{obs} = (1.0 \pm 0.2)$ km/year [50, 51]. This problem was solved by taking into account time-delay effects [51], which we review below.

1.2.2. **Time delay effects**

An important modification to Fisher's equation, that we shall use in the models in this thesis, was the introduction of a delay time [51]. As seen above (Sec. 1.1.1), in virus infections there is an 'eclipse' time between the infection of a cell and the release of the new progeny, during which neither the original virus nor the new generation diffuse. Rather similarly, when considering human populations usually the children remain with their parents until adulthood, when they move from their home to create a new family. Therefore, in both cases there is a time delay between the diffusion (or migration) of the parent and the offspring generations, which will slow down the speed of the traveling waves. For this reason, Fort and Méndez [51] introduced a time delay into the mathematical description of both systems. In mathematical terms, this corresponds to using a hyperbolic reaction-diffusion (HRD) equation, instead of the classical PRD equation (Sec. 1.2.1) that leads to Fickian diffusion and the Fisher equation (1.9). When considering a single population, an HRD equation can be derived as follows [51].

Let $p(x, y, t)$ stand for the population density at location (x, y) and time t, and τ for the time delay between two consecutive migration movements (usually one generation) [51]. If we assume that the effects of diffusion and population growth are additive (as done already below Eq. (1.7) to derive Eq. (1.1)), between times t and $t + \tau$ the processes of diffusion and reaction will cause the following change in the population density in area $ds = dx\, dy$

$$[p(x,y,t+\tau)-p(x,y,t)]ds = [p(x,y,t+\tau)-p(x,y,t)]_m ds + [p(x,y,t+\tau)-p(x,y,t)]_g ds, \quad (1.12)$$

where the subscripts m and g stand for migration and population growth processes, respectively. Let Δx and Δy stand for the spatial variations in the coordinates of a given random walk during τ (i.e., the distances between, e.g., the birthplaces of a parent and one of her/his children). Then the migration term in Eq. (1.12) can be written as

$$[p(x,y,t+\tau)-p(x,y,t)]_m ds = ds\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x+\Delta x, y+\Delta y, t)\,\phi(\Delta x,\Delta y)\,d\Delta x\,d\Delta y - ds\,p(x,y,t), \quad (1.13)$$

where $\phi(\Delta x,\Delta y)$ is the dispersion kernel, i.e. the probability per unit area that the migration distances take the values $(\Delta x,\Delta y)$. Obviously, the total probability (computed over all possible values of $(\Delta x,\Delta y)$) must add up to one, so this function satisfies that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(\Delta x,\Delta y)d\Delta x\,d\Delta y = 1$. We assume for simplicity that the low-scale migration is isotropic [98], i.e. $\phi(\Delta x,\Delta y) = \phi(-\Delta x,\Delta y) = \phi(\Delta x,-\Delta y) = \phi(\Delta y,\Delta x)$. If the increments of time and space are small enough, i.e. $\tau \ll t$, $\Delta x \ll x$ and $\Delta y \ll y$, we may Taylor-expand the new population density up to second order as

$$p(x+\Delta x,y+\Delta y,t) = p(x,y,t) + \frac{\partial p}{\partial x}\Delta x + \frac{\partial p}{\partial y}\Delta y + \frac{\partial^2 p}{\partial x^2}\frac{\Delta x^2}{2} + \frac{\partial^2 p}{\partial y^2}\frac{\Delta y^2}{2} + 2\frac{\partial^2 p}{\partial x\partial y}\frac{\Delta x\Delta y}{2}. \quad (1.14)$$

Since, as mentioned above, we assume that migration is isotropic, we have that $\phi(\Delta x,\Delta y) = \phi(-\Delta x,\Delta y) = \phi(\Delta x,-\Delta y)$ and therefore, when using Eq. (1.14) into (1.13), linear and cross-terms are suppressed and only the second space derivatives remain, i.e. $\frac{\partial^2 p}{\partial x^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(\Delta x,\Delta y)\frac{\Delta x^2}{2}\,d\Delta x\,d\Delta y$ and $\frac{\partial^2 p}{\partial y^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(\Delta x,\Delta y)\frac{\Delta y^2}{2}\,d\Delta x\,d\Delta y$.

The assumption that migration is isotropic also implies that $\phi(\Delta x,\Delta y) = \phi(\Delta y,\Delta x)$, thus $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(\Delta x,\Delta y)\frac{\Delta x^2}{2}\,d\Delta x\,d\Delta y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(\Delta x,\Delta y)\frac{\Delta y^2}{2}\,d\Delta x\,d\Delta y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(\Delta x,\Delta y)\frac{\Delta^2}{4}\,d\Delta x\,d\Delta y$, where $\Delta = \sqrt{\Delta x^2+\Delta y^2}$, and these two integrals can be written simply as $D\cdot\tau$ if we apply that $D = \frac{1}{4\tau}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\phi(\Delta x,\Delta y)\Delta^2\,d\Delta x\,d\Delta y$ is the diffusion coefficient in 2-dimensional space. This will lead to Eq. (1.16), which reduces (for $\tau\to 0$) to Eq. (1.1). This proofs that $D = \frac{\langle\Delta^2\rangle}{4\tau}$ in 2 dimensions, as mentioned below Eq. (1.1).

On the other hand, the population growth term in Eq. (1.12) can be Taylor expanded as

$$[p(x,y,t+\tau)-p(x,y,t)]_g ds = \left(\tau F(x,y,t) + \frac{\tau^2}{2}\frac{\partial F(x,y,t)}{\partial t} + \cdots\right)ds, \quad (1.15)$$

where $F(x,y,t)$ is the change in the population number due to population growth processes.

If we now introduce these results into Eq. (1.12), and Taylor-expand to second order the left-hand side of the equation, we finally achieve the HRD equation

$$\frac{\partial p}{\partial t} + \frac{\tau}{2}\frac{\partial^2 p}{\partial t^2} = D\left(\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2}\right) + F(x,y,t) + \frac{\tau^2}{2}\frac{\partial F(x,y,t)}{\partial t}. \quad (1.16)$$

This equation was first applied to describe the range expansion of the European Neolithic [51], and later to those of Paleolithic populations [110, 111], non-human species [112] and virus infections [4, 21] (in the latter case it must be supplemented with additional equations, as we shall explain in the next subsection).

If one assumes a logistic function for the reaction or population growth term, Eq. (1.8), as done above to obtain the Fisher equation, Eq. (1.16) leads to the following expression for the front speed [51]

$$c_{HRD} = \frac{2\sqrt{aD}}{1 + a\frac{\tau}{2}} \quad . \tag{1.17}$$

If $\tau = 0$, from this equation we recover the Fisher propagation speed, Eq. (1.10), and the HRD Eq. (1.16) reduces to the PRD Eq. (1.1), as it also should. For $\tau > 0$, the predicted front speed is obviously lower than the one predicted by using the Fisher equation, Eq. (1.10). This is reasonable, because the role of τ is to introduce a time delay in the diffusion process, as explained above.

Fort and Méndez applied the time-delayed speed (1.17) to obtain a spread rate that agrees better with that observed for the Neolithic in Europe than the speed due to Fisher (1.10) [51]. Similar results were obtained for Paleolithic populations [110, 111] and non-human species [112]. The time-delayed approach also yields better fits to data from *in vitro* experiments of growing virus plaques [4] than previous, non-delayed models [14]. We next review the latter application.

1.2.3. **The plaque growth problem**

Let us consider the first of the three problems that we want to address, namely the plaque growth of viral focal infections. In this subsection we review the previous mathematical models that are necessary to understand our paper on this topic (which is reproduced in Chapter 3).

The dynamics of a virus-host cell system is complex, due to intra- and extracellular interactions between invading virus particles and host cells. Nevertheless, the models summarized below try to reduce the complexity of the equations, and at the same time to account for the experimental speeds.

As mentioned in Secs. 1.1.1 and 1.2.1, Yin and McCaskill developed in 1992 the first set of reaction-diffusion equations capable of describing in a quantitative way the spatial dynamics of a system composed by viruses (V), healthy host bacteria (B) and infected cells (I) [14]. They followed previous similar work on non-viral systems [113, 114]. The relevant interactions in this system can be summarized by the reactions

$$V + B \xrightarrow{k_1} I \xrightarrow{k_2} Y \cdot V, \tag{1.18}$$

where k_1 is the rate constant of virus adsorption into uninfected cells, k_2 is the death rate of infected cells, and Y is the yield or burst size (Y is defined as the number of new viruses per lysed cell or, equivalently, per initial virus). Note that here we are dealing with three populations (V, B and I), whereas in the simpler systems considered in the previous section we had only one. For this reason,

it is not possible to find an exact, explicit speed similar to Eq. (1.17) for virus infections. However, implicit equations for the speed can be easily derived, as we next explain.

Because plaques are usually radially symmetric, Yin and McCaskill [14] proposed a set of three reaction-diffusion equations in polar coordinates. We use t to denote the time, r the distance from the inoculation point (or, equivalently, the center of the region where viruses are localized at $t = 0$), and square brackets to denote concentrations. Then, the Yin-McCaskill model is [14]

$$\frac{\partial [I](r,t)}{\partial t} = k_1 [V](r,t)[B](r,t) - k_2 [I](r,t), \tag{1.19}$$

$$\frac{\partial [V](r,t)}{\partial t} = D \frac{\partial^2 [V](r,t)}{\partial r^2} - k_1 [V](r,t)[B](r,t) + Y k_2 [I](r,t), \tag{1.20}$$

$$\frac{\partial [B](r,t)}{\partial t} = -k_1 [V](r,t)[B](r,t). \tag{1.21}$$

Because in the experiments they wanted to describe the host bacteria are immobilized in agar, only the equation describing the virus dynamics (1.20) includes a diffusive term (i.e., that depending on the diffusion coefficient D of viruses), whereas Eqs. (1.19) and (1.21) only include the terms related to the reactions (1.18). Note that Eq. (1.20) is simply a special case of Eq. (1.1).

In agreement with the first reaction in Eq. (1.18), the terms containing k_1 in Eqs. (1.19)-(1.21) imply that the concentration of infected bacteria $[I]$ increases as a result of virus infection (positive term in Eq. (1.19)), while the concentrations of free viruses $[V]$ and healthy bacteria $[B]$ decrease (negative terms in Eqs. (1.20)-(1.21)). On the other hand, the lysis process (second reaction in Eq. (1.18)) causes a decrease in the concentration of infected cells (last term in Eq. (1.19)) and an increase of free viruses (last term in Eq. (1.20)), both of them with rate k_2.

Yin and McCaskill [14] considered the boundary conditions $\frac{\partial [V](r,t)}{\partial r} = 0$ at $r = 0$ (vanishing flux), and $[V] = [I] = 0$ and $[B] = B_0$ as $r \to \infty$, and initial conditions such that all bacteria are infected within a small disk centered at $r = 0$. This makes it possible to obtain an implicit and rather complicated expression for the speed. However, it was noted that, when all parameter values are estimated from independent experiments (rather than adjusting them as in the Ref. [14]), the speeds from the model are several times faster than the observed ones [14, 19]. Later it was suggested that the eclipse or delay time (i.e. the time interval during which a virus is inside a cell and thus does not move) could affect the speed of the infection front, slowing down its spread [4]. For this reason, Fort and Méndez replaced Eq. (1.20) by the time-delayed one Eq. (1.16) to describe the dynamics of virus infections, obtaining a good fit between model and observations without fitting any parameter values [4]. After several subsequent contributions on the effect of the delay time on virus infection fronts [115, 116], Amor and Fort improved the previous approaches by using the following reaction-diffusion set of equations [21]

$$\frac{\partial [I](r,t)}{\partial t} = k_1 [V](r,t)[B](r,t) - k_2 [I](r,t) \left(1 - \frac{[I](r,t)}{I_{max}} \right), \tag{1.22}$$

$$\frac{\partial [V](r,t)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 [V](r,t)}{\partial t^2} = D\frac{\partial^2 [V](r,t)}{\partial r^2} + F(r,t) + \frac{\tau}{2}\frac{\partial [F](r,t)}{\partial t}\bigg|_g , \tag{1.23}$$

$$\frac{\partial [B](r,t)}{\partial t} = -k_1 [V](r,t)[B](r,t), \tag{1.24}$$

where the virus growth function $F(r,t)$ accounts for all reactive processes, i.e. viruses' adsorption into susceptible cells at rate k_1 (which decreases the density of free viruses) and the release of Y of viruses at rate k_2 when an infected cell dies (which increases the number of viruses),

$$F(r,t) \equiv \frac{\partial [V](r,t)}{\partial t}\bigg|_g = -k_1 [V](r,t)[B](r,t) + k_2 Y [I](r,t)\left(1 - \frac{[I](r,t)}{I_{max}}\right). \tag{1.25}$$

Note that the last parentheses in Eqs. (1.22) and (1.25) is a logistic term [such as Eq. (1.8)], and that it does not appear in Eq. (1.19). This logistic term was introduced [4] because it makes it possible to obtain agreement with experiments for homogeneous systems without adsorption, $k_1 = 0$ (these are the so-called one-step experiments). In this situation, we obtain from Eqs. (1.22)-(1.24) $\frac{d[V]}{dt} = -Y\frac{d[I]}{dt} = k_2 Y[I]\left(1 - \frac{[I]}{I_{max}}\right)$, and the solution to this equation under the appropriate boundary conditions ($lim_{[I]\to I_{max}}[V] = 0$ and $lim_{[I]\to 0}[V] = [V]_{max}$) is $[V] = \frac{Y\, I_{max}}{1+c_1 \exp(-k_2 t)}$, where c_1 is an integration constant [4]. This dependence of $[V]$ on time is consistent with the one-step data, in which $[V]$ remains fairly constant for some time, then increases steeply, and finally saturates (for details see Fig. 1 in Ref. [4]). Thus, the use of a logistic term is an improvement over Eq. (1.19), because Eq. (1.19) predicts an exponential dependence of $[V]$ on time and this is clearly inconsistent with those experiments [4]. In spite of this improvement, we shall explain a limitation of this logistic description and propose a more appropriate model in Chapter 3 [7].

In Eq. (1.23), which is the same as Eq. (1.16), the symbol $\ldots|_g$ indicates that time derivatives of F are related exclusively to the reactive process (and not to the diffusive one) [117]. This point, namely the proper computation of the terms with the symbol $\ldots|_g$ in Eq. (1.23), is the main theoretical improvement of the model by Amor and Fort [21] relative the original time-delayed model [4]. Note that the Yin-McCaskill Eqs. (1.19)-(1.21) are not time-delayed, i.e. they do not include the first and last terms that appear in Eq. (1.23) (terms of this kind were first included in virus infections in Ref. [4], and we stress that they are the same as the corresponding ones in Eq. (1.16)). The time derivative of the reactive function F (which appears in Eq. (1.23)) is

$$\begin{aligned}\frac{\tau}{2}\frac{\partial F(r,t)}{\partial t}\bigg|_g &= -\frac{\tau}{2}k_1 \frac{\partial\{[V](r,t)[B](r,t)\}}{\partial t}\bigg|_g + \frac{\tau}{2}k_2 Y\frac{\partial}{\partial t}\left[[I](r,t)\left(1 - \frac{[I](r,t)}{I_{max}}\right)\right] \\ &= -\frac{\tau}{2}k_1 [V](r,t)\frac{\partial [B](r,t)}{\partial t} - \frac{\tau}{2}k_1 [B](r,t)F(r,t) \\ &\quad + \frac{\tau}{2}k_2 Y\frac{\partial}{\partial t}\left[[I](r,t)\left(1 - \frac{[I](r,t)}{I_{max}}\right)\right].\end{aligned} \tag{1.26}$$

Amor and Fort [21] obtained an analytical, implicit solution for the speed, and checked it by integrating numerically the model (1.22)-(1.26) using parameters values obtained from independent experiments [21]. The simulations agree with the observed data, and the mathematical description is sounder than that in the first model by Yin and McCaskill [14]. Still, as mentioned above, the use of a

logistic function in Eq. (1.22) is questionable from a biological point of view. We shall explain this problem in Chapter 3, and use this mathematical model as our starting point to develop a biologically sounder model. Amor and Fort were also the first to apply the time-delayed approach to VSV infections [21] (previous work had used experimental data for the T7-*E. Coli* system). VSV is the virus that we have studied for its oncolytic effect in Chapter 4.

1.2.4. **Oncolytic treatment of cancer tumors**

The second problem addressed in this thesis is related to virus treatment of cancer tumors. Due to advances in genetic engineering, medical interest in oncolytic virotherapy has been renewed recently [118, 119, 120, 30, 31]. In turn, this has stimulated novel mathematical models to describe oncolytic processes [40, 46, 121]. Full understanding of virus-tumor dynamics is still far away from us, partly because of the complexity of the genetics of the tumors themselves [122, 123, 124]. Nonetheless, mathematical and computational modelling, as well as statistical analysis of macroscopic data, have contributed to improve our understanding of some aspects of tumor [125, 34] and oncolytic [126, 127, 128, 45, 121] dynamics. In spite of the fact that we do not understand the full complexity of the process, mathematical and computational modelling can describe important processes affecting treatments with oncolytic viruses.

Many authors have developed mathematical models to describe of oncolytic systems (see Sec. 0). Because we will analyze front speeds (Chapter 4), we are especially interested in spatial models [41, 42, 43, 44, 45, 46]. In this subsection we introduce the model by Wodarz et al. [45, 46], and later we will improve it (Chapter 4). In Ref. [45], Wodarz et al. compared the results from a mathematical model, based on a set of reaction-diffusion partial differential equations (PDEs), as well as those from an agent-based computational model, to the results obtained from *in vitro* experiments performed with a newly constructed virus. Here we shall focus on their PDE model. The main advantage of PDE models (as compared to agent-based models) is that they usually allow to find front speeds as a function of parameter values fast (compared, e.g., to agent-based models, in which it is necessary to repeat many simulations, which takes substantially more computing time). In their model, Wodarz et al. [45] consider the dynamics of only two populations, uninfected tumoral cells T and infected tumoral cells I. In order to derive their model, consider first the following two equations

$$\frac{\partial [T](r,t)}{\partial t} = D_T \frac{\partial^2 [T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[I](r,t) + [T](r,t)}{k}\right\} - k_1[V](r,t)[T](r,t), \quad (1.27)$$

$$\frac{\partial [I](r,t)}{\partial t} = D_I \frac{\partial^2 [I](r,t)}{\partial r^2} - k_2[I](r,t) + k_1[V](r,t)[T](r,t). \quad (1.28)$$

where, as in Eqs. (1.19) and (1.21), parameters k_1 and k_2 are the rate of adsorption of viruses into uninfected tumor cells and the rate of death for infected tumor cells, respectively. Parameter a is the proliferation rate of the uninfected tumor cells T, and k is the local carrying capacity of tumor (infected and non-infected) cells. Thus, the term containing the symbols { } in Eq. (1.27) is just an example of logistic growth, Eq. (1.8). Note that in Eq. (1.21) this term did not appear because at the beginning of that experiment the cell population is already saturated all over that system, and the reproduction of cells is very slow compared to their death due to infection. Similarly, in the tumor-virus system considered in this subsection, if most of the cells infected by the virus die before reproducing, a term describing the proliferation rate of infected cells (analogous to the second term in the right-hand side

of Eq. (1.27)) can be neglected in Eq. (1.28), as done by Wodarz et al. [45]. The first term on the right-hand side of both equations describes the diffusion of the T and I cells, which in general are ruled by their own diffusion coefficients, D_T and D_I, respectively (recall than in Eqs. (1.19) and (1.20) there are no diffusive terms because the cells are immobilized by agar in those experiments). Note that in Eqs. (1.27)-(1.28) we have three concentrations ($[T]$, $[I]$ and $[V]$), so a third equation for the virus dynamics (e.g., Eq. (1.20) with $[B]$ replaced by $[T]$) is necessary to find the evolution of $[T]$, $[I]$ and $[V]$ in space and time. However, as explained in detail below, Wodarz et al. [45] considered the dynamics of only two concentrations.

Wodarz et al. [45] took Eqs. (1.27)-(1.28) from a previous homogeneous model (i.e., $\frac{\partial^2[\dots]}{\partial r^2} = 0$) by Nowak and May [35], which had originally three equations with three variables (viruses V, susceptible cells T and infected cells I), similarly to Eqs. (1.19)-(1.21). Nowak and May [35] wrote down Eq. (1.20) including a rate k_3 of decay for free viruses (natural death) in homogeneous systems ($\frac{\partial^2[V](r,t)}{\partial r^2} = 0$) as

$$\frac{\partial[V](r,t)}{\partial t} = Yk_2[I](r,t) - k_3\,[V](r,t). \tag{1.29}$$

Nowak and May [35] neglected the term $-k_1[V](r,t)[B](r,t)$ in Eq. (1.20), because in some systems this turns out to be a good approximation. Usually we include this term in our models (but for the T7-*E. Coli* system it can be neglected, and the results become simpler, as discussed in Sec. 3.5 in this thesis).

There are two problems with the model due to Wodarz et al. [45]. The first problem is that, following Nowak and May [35] , they assume that free viruses are approximately in a steady state, i.e. $\frac{\partial[V]}{\partial t} \approx 0$ and therefore Eq. (1.29) yields $[V](r,t) \approx \frac{k_2 Y}{k_3}[I](r,t)$ [35, 45]. Wodarz et al. used [45] this special relationship between $[V]$ and $[I]$ to write down their mathematical model from Eqs. (1.27)-(1.28) as

$$\frac{\partial[T](r,t)}{\partial t} = D_T\frac{\partial^2[T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[I](r,t)+[T](r,t)}{k}\right\} - b[I](r,t)[T](r,t), \tag{1.30}$$

$$\frac{\partial[I](r,t)}{\partial t} = D_I\frac{\partial^2[I](r,t)}{\partial r^2} - k_2[I](r,t) + b[I](r,t)[T](r,t). \tag{1.31}$$

where $b = \frac{k_1 k_2 Y}{k_3}$. Note that, in contrast to Eqs. (1.27)-(1.28), there are only two concentrations in Eqs. (1.30)-(1.31), namely $[T]$ and $[I]$.

Nowak and May [35] argue that the virus quasi-steady approximation (i.e., $\frac{\partial[V]}{\partial t} \approx 0$) is valid if $k_3 \gg k_2$, i.e. if viruses die much faster than the infected cells (see the last term in Eqs. (1.29) and (1.19)). The same special case ($k_3 \gg k_2$ and, therefore, $\frac{\partial[V]}{\partial t} \approx 0$) had been introduced previously (in a different context) by May and Anderson [129], who justified it by arguing that $k_3 \gg k_2$ implies that the virus concentration decays so fast that it becomes adjusted essentially instantaneously, for any given values of $[T](r,t)$ and $[I](r,t)$, to its local equilibrium level, namely $[V](r,t) = \frac{k_2 Y}{k_3}[I](r,t)$ (from Eq. (1.29) with $\frac{\partial[V]}{\partial t} = 0$). Thus, in case the virus concentration surpasses this local equilibrium

value, it will diminish very rapidly until virus deaths (last term in Eq. (1.29)) are compensated by the production of new viruses (last-but-one term in Eq. (1.29)). Therefore, this production term avoids the total disappearance of viruses.

In contrast to Wodarz et al. [45], we will not apply the quasi-steady approximation to oncolytic systems because, as argued in the next paragraph, neither $k_3 \gg k_2$ nor $\frac{\partial[V]}{\partial t} \approx 0$ are satisfied in the oncolytic systems that we analyze in Chapter 4.

Using realistic parameter values of k_3 and k_2, we can check whether the assumption $k_3 \gg k_2$ is valid or not. For the system that we will analyze in Chapter 4, these parameters will be estimated in Secs. 2.1.3 and 2.1.4 from experimental data. It turns out that their numerical ranges (namely, $0.017 < k_2 < 0.042\ h^{-1}$ and $0.014 < k_3 < 0.028$ h^{-1}) are of the same order of magnitude. This shows conclusively that the assumption $k_3 \gg k_2$ is not valid for our purposes. Similarly, for the non-oncolytic virus system discussed in Chapter 3, $k_3 \approx 0$ and, again, the assumption $k_3 \gg k_2$ canot be made.

Moreover, when a travelling wave of viruses propagates, it is easy to argue that the assumption $\frac{\partial[V]}{\partial t} \approx 0$ is not justified. An intuitive way to see this is the following. Outside the initially infected area, before the arrival of the infection, there are no viruses ($[V] = 0$), when the infection arrives $[V]$ increases ($\frac{\partial V}{\partial t} > 0$), and after cells are killed $[V]$ decreases again ($\frac{\partial V}{\partial t} < 0$). This suggests that the condition $\frac{\partial V}{\partial t} \approx 0$ cannot be assumed for the systems that we analyze in this thesis.

Before dealing with the second problem, we mention that other authors, e.g. [130], justify the quasi-steady approximation $\frac{\partial[V]}{\partial t} \approx 0$ in a different way, namely by assuming that viruses reproduce much faster than tumor cells ($Yk_2 \gg a$). Here the idea is that a long time is necessary for a new cell to appear (compared to the time spent by a virus to reproduce inside an infected cell). This means that, after a virus reproduces, its progeny finds almost no cells to infect for a long time, during which $\frac{\partial[V]}{\partial t} \approx 0$ is a realistic approximation. In homogeneous systems such that $Yk_2 \gg a$ this is reasonable, but in this thesis, we do not deal with homogeneous systems. The systems considered by us are clearly non-homogeneous, because in the infected region (red area and curve in Fig. 1.1) there are only free viruses, in the infection leading edge there are infected cells and free viruses (purple and red curves in Fig. 1.1), and in the non-infected region there are only non-infected cells (blue region a curve in Fig. 1.1). In such systems, when a virus front arrives to a region with uninfected cells, it obviously finds many cells to infect, even if they do not reproduce at all.

For all of the reasons above, in our opinion the assumption $\frac{\partial[V]}{\partial t} \approx 0$ is not justified (even if $Yk_2 \gg a$). So, in contrast to Wodarz et al. [45], we will not make use of this assumption.

A second problem with the model by Wodarz et al. [45] (Eqs.(1.30)-(1.31)) is that it does not take into account the delay time (see the text above Eq. (1.22)), but we know from previous work [4, 21] that its effect can be very important.

In conclusion, a different model is necessary for our purposes, and we will introduce it in Chapter 4.

It is worth to mention that Wodarz et al. [45] noted the following interesting application of the study of front speeds in virus treatments of cancer tumors. Recall that, for virus infections, if initially only

the cells of a small region are infected (this would correspond to a very small area in the center of the red region in Fig. 1.1), we can distinguish three well-defined regions, namely (i) an inner area containing infected cells ($I \neq 0$), corresponding to the red region in Fig. 1.1, surrounded by (ii) a region where host cells are at their maximum carrying capacity and which has not yet been reached by the infection front ($T = k$ and $I = 0$, blue region in Fig. 1.1) and (iii) the outermost medium without presence of any cell ($T = I = 0$, white region in Fig. 1.1). The virus front advances from region (i) outwards into region (ii), and we will denote its speed by c_I. The tumor front advances from region (ii) outwards into region (iii), and we will denote its speed by c_T. From Eqs. (1.30)-(1.31), these speeds can be estimated as [45] $c_I = 2\sqrt{D_I(k_1 k - k_2)}$ and $c_T = 2\sqrt{D_T a}$. Note that $c_T = 2\sqrt{D_T a}$ is nothing but the Fisher speed (Eq. (1.10)), because for the tumor front we are dealing simply with an expanding population of uninfected tumor cells. On the other hand, $c_I = 2\sqrt{D_I(k_1 k - k_2)}$ can be obtained from Eq. (1.31) by noting that, in the leading edge of the virus front, $[T] \approx k$ and comparing to the Skellman equation, Eq. (1.11). The important point is that, if $c_I > c_T$, the viruses can eventually kill the tumor, whereas otherwise they will never reach all of the tumor cells [45]. We caution that the result $c_I = 2\sqrt{D_I(k_1 k - k_2)}$ by Wodarz et al. [45] will break down for the system we are interested in (due to the two problems explained above). However, this point highlights the usefulness of comparing the values of c_I and c_T, and thus the importance of the study of the speeds of advancing waves in virus treatments of cancer tumors.

1.2.5. **Neolithic spread and human interaction**

The third application studied in this thesis is related to the geographic expansion of the Neolithic across Europe. In this thesis we use the term 'Neolithic' to denote farming and stockbreeding, as usually done by most archaeologists in Western Europe (in contrast, in Russia and Eastern Europe sometimes 'Neolithic' denotes the use of pottery, not necessarily by farmers). During the early and mid-twentieth century, qualitative analysis of archaeological data led to the clear conclusion that European farming originated in the Near East, from where it spread gradually across Europe [131, 132, 133, 134, 135, 136]. As we have mentioned above (end of Sec. 1.2.1), Ammerman and Cavalli-Sforza were the first to apply a statistically sound analysis to the archaeological data [50], and to apply the Fisher equation (1.9) to describe the spread of early farmers across Europe [53]. Since then, many authors have developed new mathematical and computational models to better describe the dynamics of the Neolithic spread at continental and local scales [92, 137, 3, 95, 138, 48]. As one example of such models, above we have discussed the effect of a delay time (see Eq. (1.16)) related to the fact that newborn humans spend some time with their parents before leaving them [51].

However, while differential-equation models such as the Fisher equation (1.9) or the HRD equation (1.16) are useful to predict average behaviors, they cannot capture effects due to a real, non-homogenous geography. For this reason in 2012 Fort, Pujol and Vander Linden [47] developed a computational model to describe the Neolithic transition in which they took into account the effects of sea travel and mountain barriers. Their simulation runs on a rectangular grid of 180×102 square cells covering the whole European continent and part of the Near East. Each cell is a square with side equal to 50 km, because this is the value corresponding to the mobility per generation obtained from measured data for preindustrial farmers [52, 139]. The Neolithic population is initially present only at some regions in the Near East, specifically where pre-pottery Neolithic B/C (PPNB/C) sites have been found. The reason is that the formation of the Neolithic in the Near East was not a front propagation

phenomenon at the beginning, because some innovations appeared earlier, others later on, and in different areas [140]. A really homogeneous package of domestic plants and animals formed only later, with the PPNB/C cultures, and it was this well-defined cultural package that later spread across Europe [141]. The PPNB/C sites in the database used by Fort, Pujol and Vander Linden [47] are located in Israel, Jordan, Lebanon, Iraq, Syria and Turkey. In the numerical simulation model by Fort, Pujol and Vander Linden [47], the Neolithic spread by following a two-step reaction-diffusion scheme that is repeated at each generation. Firstly, the new number of Neolithic farmers (P_F) due to population growth is computed in each cell and at each time step (generation) according to the following rules:

$$\begin{aligned} {P'}_F(i,j,t) &= R_{0,F}\, P_F(i,j,t) \quad \text{if} \quad P_F(i,j,t) < {P_{F\,max}}\big/{R_{0,F}} \\ {P'}_F(i,j,t) &= P_{F\,max} \quad \text{if} \quad P_F(i,j,t) \geq {P_{F\,max}}\big/{R_{0,F}} \end{aligned}, \tag{1.32}$$

where $R_{0,F}$ is the net reproductive rate (or fecundity) per generation[1] and $P_{F\,max}$ corresponds to the maximum sustainable population in a cell. Alternatively, a logistic function could be also used, but the results would not change. Secondly, the effect of dispersion is computed at each time step. In the case of homogeneous land travel, the simple model used in Ref. [47] assumes that each generation time T, a fraction p_e (persistence) of the initial population in each cell will not disperse and one fourth of the remaining population will move to one of the four nearest neighbor cells (i.e., with center 50 km away from that of the original cell). This process can be described by the following equation:

$$\begin{aligned} P_F(i,j,t+1) = {} & p_e\, {P'}_F(i,j,t) \\ & + \frac{1-p_e}{4}\left[{P'}_F(i-1,j,t) + {P'}_F(i+1,j,t) + {P'}_F(i,j-1,t)\right. \\ & \left. + {P'}_F(i,j+1,t)\right], \end{aligned} \tag{1.33}$$

where i and j are the grid coordinates of the original cell. Alternatively, more complicated dispersal models involving several distances could be used, but the authors of Ref. [47] expect that the main conclusions would be the same and prefer to focus on non-homogenous effects. When mountain and sea cells are considered, the dispersal process is no longer homogeneous. If one of the four neighboring cells is a mountain cell, farmers leaving the original cell will redistribute among the remaining three cells. On the other hand, if one of the four neighboring cells is a sea cell, the population that would move there is assumed to travel by sea to coast cells within a given range. A more detailed explanation of this type of computational model is given in Sec. 2.3.3 and Chapter 5 of this thesis, because we use the same kind of simulations to model the genetic consequences of the Neolithic spread (in contrast, no genetics simulations were performed in Ref. [47]). Note, however, that there is a direct analogy between the computational model described here and the differential-equation models in Eqs. (1.9) and (1.16). In addition, the time delay effect from the HRD equation (1.16) is implicitly included in this kind of simulations. In their simulations, Fort, Pujol and Vander Linden [47] noted that mountains do not hinder the expansion remarkably. On the other hand, assuming travel distances of up to 150 km and using parameter values from ethnographic data, they obtained good agreement between their simulations and a database of 919 dated sites (at the

[1] $R_{0,F}$ is directly related to the growth rate a (introduced in Sec. 1.2.1) as $a = \frac{\ln R_{0,F}}{T}$. The next chapter (Materials and methods) describes in detail all of the parameters used in this thesis; for more information about this parameter see Sec. 2.2.6.

continental level). They observed that, in contrast to mountains, sea travel had an important effect, by accelerating the Neolithic spread along the Mediterranean.

Some authors have also performed space-time genetic simulations of the Neolithic spread. They did not compare to ancient DNA data, because they were then unavailable (the first comparison to the shape of an ancient genetic marker is presented in Chapter 5 in this thesis). Let us summarize two of these approaches [5, 89]. In 1986, Rendine et al. published one of the earliest such simulations [5]. Their computational model is similar to that in Ref. [47] described above, but also includes a cultural interaction process, according to which in each step, following a Lotka-Volterra approach, $\gamma P_{HG}(t)P_F(t)$ hunter-gatherers become farmers at each cell (with $P_{HG}(t)$ and $P_F(t)$ the initial number of hunter-gatherers and farmers at the cell considered, respectively). Also, the population frequencies of several genes were computed, in order to see how genetic gradients (which also called genetic clines) form, depending on the initial frequencies. Additionally, by assuming several migrations (from different regions and times), it was shown that the effects of independent dispersals can be recognized by means of a statistical method called principal components [5] (which had been previously used to analyze the genetics of present European populations [54]).

In another spatial simulation study on the genetic consequences of the Neolithic transition, Currat and Excoffier [89] studied different degrees of demic and cultural diffusion. Their simulation is inspired by that from Rendine et al. [5]. One difference between both studies is that by Currat and Excoffier [89] include, besides the spread of Neolithic farmers from the Near East, the first arrival and spread of Paleolithic populations (modern humans) across Europe, after their out-of-Africa dispersal. Another difference between both studies is that the acculturation process considered by Currat and Excoffier [89] is described mathematically as $\gamma \frac{2P_{HG}P_F}{(P_F+P_{HG})^2}$, where $\gamma = 0$ corresponds to a fully demic diffusion model. Currat and Excoffier [89] found that minute amounts of cultural diffusion have important effects on the genetic composition of the farming population. They also found that clines with extreme frequencies in the Near East could have formed not only due to the Neolithic spread, but also due to the arrival and spread of the first modern humans into Europe. The reason of the latter clines is a series of founder effects at the population wavefront, in which specific markers increase their frequencies because of random effects due to the low population density [89].

Although the two previous models [5, 89] include cultural terms, none of them are based on cultural transmission theory [93]. This theory formalizes mathematically the three different kinds of cultural transmission, namely vertical transmission (which corresponds to the transmission of cultural traits from parents to their children), horizontal transmission (from some individuals to others of the same generation), and oblique transmission (from some individuals of a generation to others of the next one, excluding their children) [93]. Recently, vertical cultural transmission effects on Neolithic spread where analyzed by Fort [94]. He used cultural transmission theory [93] to describe the effects of cross-matings between hunter-gatherers (HGs) and farmers (Fs) on Neolithic spread. In agreement with ethnographic data [142, 143], it can be assumed that when farmers mate with hunter-gatherers their children will be all farmers. Thus, the number of farmers P_F increases and the number of hunter-gatherers P_{HG} decreases as a result of this interaction. In this framework, Fort [94] derived the following reaction terms (which include growth and interaction) for farmers (F) and hunter-gatherers (HG)

$$P_F(t+T) = R_{0,F}P_F(t) + R_{0,F}\eta\frac{P_F(t)P_{HG}(t)}{P_F(t)+P_{HG}(t)}, \quad (1.34)$$

$$P_{HG}(t+T) = R_{0,HG}P_{HG}(t) - R_{0,HG}\eta\frac{P_F(t)P_{HG}(t)}{P_F(t)+P_{HG}(t)}, \quad (1.35)$$

where T is the generation time, and η is the intensity of interbreeding. A simplified derivation of Eqs. (1.34)-(1.35) is the following. Let $u = \frac{P_F(t)}{P_F(t)+P_{HG}(t)}$ stand for the frequency of farmers, and $p'(u)$ stand for the probability that a HG mates a farmer, i.e. the number of cross-matings divided by $P_{HG}(t)$. Then $P_F(t+T) = R_{0,F}P_F(t) + I_N$, where $I_N = R_{0,F}\,p'(u)P_{HG}(t)$ is the interaction term due to interbreeding. Now, if we assumed that $p'(u)$ is independent of u, we would obtain two unacceptable results: (i) I_N would not increase with increasing values of $P_F(t)$; and (ii) $I_N \to \infty$ for $P_{HG}(t) \to \infty$, which again is not reasonable because we expect that, if $P_{HG} \gg P_F$, then HGs will have reached their maximum number of social contacts (encounters per unit time, personal relationships, etc.) with farmers, so that I_N should not diverge but rather saturate at a finite value. On the other hand, if we assumed that $p'(u)$ were proportional to u^2, or u^3, or u^4, etc. (or a linear combination of such powers), then we would obtain $I_N \to 0$ if $P_{HG}(t) \to \infty$, which is obviously unreasonable. Finally, if $p'(u)$ is proportional to u, say $p'(u) = \eta u$, then these problems do not arise and we obtain Eqs. (1.34)-(1.35). A more detailed derivation and discussions can be found in Ref. [94]. That work also combines Eqs. (1.34)-(1.35) with a reaction-diffusion model and, by comparing the model predictions to the average continental speed of the Neolithic front in Europe, concludes that the interbreeding parameter would have been $\eta < 0.1$, implying a relatively low importance of cultural diffusion against demic diffusion. We will use this model in Chapter 5 [9]. On the other hand, Eqs. (1.34)-(1.35) have been also applied in another very recent model, albeit focused in a specific region (the Western Mediterranean) and without any genetic analysis [48].

Some previous spatial genetic simulations have been based on Lotka-Volterra equations [5], i.e. on replacing Eqs. (1.34)-(1.35) by $P_F(t+T) = R_{0,F}[P_F(t) + \gamma P_{HG}(t)P_F(t)]$ and $P_{HG}(t+T) = R_{0,HG}[P_{HG}(t) - \gamma P_{HG}(t)P_F(t)]$. But we can see that this is problematic if we note, e.g., that problem (ii) mentioned below Eq. (1.34) also arises for Lotka-Volterra equations.

In 2012, Fort [3] proposed a model to describe horizontal/oblique cultural transmission effects on Neolithic spread, again based on cultural transmission theory [93]. Thus, this model describes the adoption of farming by hunter-gatherers which is not due to interbreeding but to the learning of agricultural techniques from the farmers of the same (horizontal) or the previous (oblique) generation. Assuming again that the transition takes place only towards learning farming (not hunting and gathering), as a result of the interaction the number of farmers P_F increases and the number of hunter-gatherers P_{HG} decreases as [3]

$$P_F(t+T) = R_{0,F}P_F(t) + R_{0,F}f\frac{P_F(t)P_{HG}(t)}{P_F(t)+\gamma P_{HG}(t)}, \quad (1.36)$$

$$P_{HG}(t+T) = R_{0,HG}P_{HG}(t) - R_{0,HG}f\frac{P_F(t)P_{HG}(t)}{P_F(t)+\gamma P_{HG}(t)}, \quad (1.37)$$

where f can be called the intensity of cultural transmission (note that the cultural effect vanishes if $f = 0$), and γ measures the preference of hunter-gatherers to copy the behavior of farmers (rather

than hunter-gatherers) if $\gamma < 1$ (or, on the contrary, to copy other hunter-gatherers if $\gamma > 1$). The derivation of Eqs. (1.36)-(1.37) is not included here (because it is longer than that of Eqs. (1.34)-(1.35)) but can be found in Ref. [3]. That work also combines Eqs. (1.36)-(1.37) with a demic model, and by comparing the result to the speed from archaeological data concludes that demic diffusion played a more important role in the spread of the Neolithic than cultural diffusion. It is worth noticing that for random copying ($\gamma = 1$), the previous Eqs. (1.36)-(1.37) would be completely analogous to Eqs. (1.34)-(1.35), with the acculturation parameter *f* in Eqs. (1.36)-(1.37) playing the same mathematical role as the interbreeding parameter η in Eqs. (1.34)-(1.35).

As mentioned above, some previous work has used Lotka-Volterra equations [5]. A very important difference between them and our acculturation Eqs. (1.36)-(1.37) can be seen by considering, e.g., the simple case $R_{0,HG} = 1$ (no net population growth) and Eq. (1.37). Then, the number of hunter-gatherers converted per farmer, namely $\frac{P_{HG}(t+T)-P_{HG}(t)}{P_F(t)}$, is equal to $f\frac{P_{HG}(t)}{P_F(t)+\gamma P_{HG}(t)}$, and this has an upper bound for increasing values of $P_{HG}(t)$, namely $\frac{f}{\gamma}$. This saturation effect is absent in the Lotka-Volterra approach, because then Eq. (1.37) is replaced by $P_{HG}(t+T) = P_{HG}(t) - \gamma P_{HG}(t)P_F(t)$, thus the number of hunter-gatherers converted per farmer, $\frac{P_{HG}(t+T)-P_{HG}(t)}{P_F(t)}$, is equal to $\gamma P_{HG}(t)$ and this increases withouth bound for increasing values of $P_{HG}(t)$. The latter result is not reasonable because obviously, the number of HGs that a farmer can convert during his lifetime cannot be arbitrarily large [3].

1.3. Models in this thesis

In this thesis we have developed new reaction-diffusion models based on the models summarized in Sec. 1.2. In the rest of this Chapter, we provide an overview of the most important features of the models presented in this thesis (Chapters 3-5).

1.3.1. Plaque growth models with more biological sense

Previous plaque growth models (Sec. 1.2.3) present several drawbacks to correctly describe virus infection dynamics.

First, the reaction-diffusion model by Yin and co-workers [14, 144] assumes Fickian diffusion with no delay time. However, as noted by Fort and Méndez, neglecting the delay time as in Refs. [14, 144] predicts speeds up to an order of magnitude faster than those observed in *in vitro* experiments (Fig. 2 in Ref. [4]). As explained in Sec. 1.2.3, the cause for this difference between predictions and experiments is that the time during which viruses replicate within a cell (and thus do not diffuse) reduces the effective speed of the infection front. Therefore, in our model in Chapter 3 we will use an HRD equation (similar to Eq. (1.16)), which includes the delay time, so that we can properly describe the dynamics of the virus front.

Equations (1.22)-(1.26) correspond to the model introduced by Amor and Fort [21]. That model and similar ones do already incorporate the delay time in the virus diffusive process (either with a single delay time [4, 115, 21, 145] or with a distribution of delay times [116]). However, in all of those models a logistic function is used to describe the growth dynamics of viruses and the decay of infected cells [positive term in Eq. (1.25) and negative term in Eq. (1.22), respectively]. This is biologically

questionable because, according to this assumption, the number of infected cells that die per unit time is given by $-k_2[I](r,t)\left(1-\frac{[I](r,t)}{[I]_{max}}\right)$, which implies that infected cells die at a rate k_2 proportionally to the density of infected cells $[I](r,t)$, but also to the 'free' space $1-\frac{[I](r,t)}{[I]_{max}}$ (i.e., the space not occupied by infected cells). The latter point is strange from a biological perspective. It is true that such a dependency is assumed in logistic growth (Eq. (1.8)), but that refers to a net reproduction process. It is biologically reasonable that such a rate can be proportional to the free space because this can happen, for example, if the net reproduction rate is proportional to the nutrients available and those are in turn proportional to the free space. But here we are dealing with a purely death process, and there is no intuitive reason why such a rate could be proportional to the free space. In contrast, our new model (Chapter 3) uses the more reasonable assumption that the number of cells dying in a given instant is proportional to the number of infected cells at some previous instant. If there is a delay time τ between virus adsorption and the beginning of the release of the new progeny, then it is reasonable to describe the death of infected cells as $-k_2[I](r,t-\tau)$, i.e. proportional to the density of infected cells at a time $t-\tau$. Note that this delay-time formulation does not affect the diffusive process (in contrast to Eqs. (1.16) or (1.23)) but a reactive one (the death of infected cells). Naturally, because each dead cell releases Y viruses, this correction also affects the virus growth dynamics.

A reactive delay time in virus infections has been already considered by other authors [146, 147], albeit with two important differences with our work. The first one is that they do not assume the infected death rate $-k_2[I](r,t-\tau)$ but $-k_1[V](r,t-\tau)[B](r,t-\tau)$ (with an additional factor in Ref. [146]). Therefore, their basic equation for the infected cell concentration is $\frac{\partial[I](r,t)}{\partial t}=k_1[V](r,t)[B](r,t)-k_1[V](r,t-\tau)[B](r,t-\tau)$. Note than both terms have the same rate k_1 (in contrast to, e.g. Eq. (1.19)), thus they assume that all cells die at the same time after infection. This disagrees with experimental data (see Chapter 3). The other difference with our work is that they do not take into account the effect of the delay time on the diffusive process. In Chapter 3 we also compute front speeds using those previous models [146, 147].

In Chapter 3, we attempt to develop a more realistic approach to the dynamics of growing plaques than all of those previous works. For this purpose, we consider the effects of the time delay both in the diffusive process as well as in the reactive ones (infected cell death and virus growth). In Chapter 3 we also develop an approximation to our new model, which yields realistic results and avoids cumbersome mathematical equations.

1.3.2. **The crucial delay time in oncolytic viral assays**

In Sec. 1.1.2 we have described the current interest on the possibility of treating some deadly brain tumors called glioblastomas (GBMs) with virus infections. Chapter 4 presents three increasingly realistic mathematical models developed to describe *in vitro* experiments on the oncolytic effect of VSVs on GBMs. Our approach avoids several problems of the model by Wodarz et al. [45], which we have introduced in Sec. 1.2.4 [Eqs. (1.27)-(1.28)].

As already mentioned in Sec. 1.2.4, the assumption of a steady-state for free virus (so that $[V]\propto[I]$), which allowed Wodarz et al. [45] to use a simplified two-equation model, does not hold for the

VSV-GMB system that we want to study (see Chapter 4 for more details). For this reason, in our first approach to modeling the virus-tumor dynamics (model 1), we will adapt the model by Wodarz et al. [45] to a three-equation system. Similarly, to the original model by Wodarz et al. [45], we consider Fickian diffusion in our model 1. Thus, the first oncolytic model developed in this thesis is similar to Yin and McCaskill's model [14] to describe a growing plaque, Eqs. (1.19)-(1.21), but including tumor growth and the decay rate of viruses k_3.

As we have argued above and has been shown in previous works [4, 21], the time delay effect (which results from the eclipse time τ elapsed between cell infection and the release of the new progeny) plays a very important role in virus infection dynamics. Therefore, it is clearly important to take it into account in order to describe the virus-tumor dynamics in oncolytic treatments. For this reason, our models 2 and 3 in Chapter 4 include delay-time effects. In order to distinguish the importance of each effect, model 2 includes the delay time only into the reactive process (i.e., the death of infected cells and the growth of the virus population), using the term $-k_2[I](r, t-\tau)$, which we have introduced in the previous subsection. Finally, in Chapter 4 we also incorporate the effect of the delay time on the diffusive dynamics of viruses into our model 3. Therefore, in Chapter 4 we study individually the effect of incorporating the time delay into the different processes and find that only when it is included in both the reactive and diffusive dynamics, can we achieve the best agreement with the experimental data. Also, the sensitivity analysis performed in Chapter 4 allows us to determine that the most important parameter in determining the rate of spread of the infection front of an oncolytic virus (and thus related to the effectiveness of the virus in defeating a growing tumor) is precisely the eclipse time or delay time τ, whereas other parameters, such as the rate of adsorption k_1 or the burst size Y, have little effect on the front speed.

1.3.3. **Analysis and modeling of ancient clines of mitochondrial DNA**

As explained in Sec. 1.1.3, the genetic data of modern populations are highly affected by population movements that have taken place after the Neolithic transition. For this reason, in order to analyze the genetic consequences of the Neolithic spread, it would be of great interest to identify genetic clines (i.e., spatial variations in the frequency of genetic markers) from ancient DNA (aDNA) data. Fortunately, during the last decade it has become possible to extract DNA data from ancient individuals. We have used all such published data for the Early and Middle European Neolithic to assemble a database (included as Appendix A Data S1). It contains all of the published genetic information on Early and Middle Neolithic European individuals, at the time of writing the paper reproduced in Chapter 5 (i.e., up to August 2017). As explained in Sec. 1.1.3, The individual sequences of Y-DNA and mtDNA are known as haplotypes, but similar haplotypes with a common ancestor are usually grouped under haplogroups.

In Chapter 5, we use the database (Appendix A Data S1) to identify a very clear cline for the frequency of mitochondrial haplogroup K among the early Neolithic farmers. This frequency decreases with increasing distance from the Near East (we stress that, as explained at the end of Sec. 1.1.3, not all haplogroups are expected to display such a clear cline, because different haplogroups can be affected by different processes such as selection and drift). We use this cline to infer information on the primacy of demic or cultural transmission processes in the Neolithic spread in Europe. Indeed, in Chapter 5 the percentage of farmers involved in cultural transmission processes is estimated by comparing the observed cline to those predicted by a demic-cultural model. In order to obtain

simulated genetic clines, we have developed a reaction-diffusion computational model based on the model by Fort, Pujol and vander Linden [47] described in Sec. 1.2.5. Our simulations run on a grid of square cells encompassing the entire European continent and part of the Near East, from where the Neolithic spreads to the whole continent. Similarly to the model by Fort, Pujol and vander Linden [47], we use a realistic geography. Thus, Neolithic individuals can move by land or by sea, and mountains act as barriers. Nonetheless, the model in Chapter 5 includes two important features which were not taken into account in Ref. [47] (because that work had different aims), namely cultural transmission and genetics.

In the model in Chapter 5 we will define three populations: farmers who have the Neolithic marker K, farmers who do not have the Neolithic marker K, and hunter-gatherers (HG). None of the HGs has the Neolithic marker K, in agreement with observations (Sec. 5.8.2). Cultural transmission between the farmer and HG populations present in each cell will be taken into account by assuming a process of cultural transmission ruled by Eqs. (1.34)-(1.35) [94]. In principle, we could use Eqs. (1.36)-(1.37) instead, but those equations have two unknown parameters and, in any case, we expect that the final conclusions would be much the same (Sec. 5.8.9). Cultural transmission has been also included in a very recent model of the Neolithic spread in the Western Mediterranean [48], but the model in Chapter 5 is more complex because of the three population groups considered. This makes it possible (in contrast to Ref. [48]) to analyze the genetics, which was not done in Ref. [48]. The number of individuals in each generation is calculated by means of growth processes of the type described by Eq. (1.31); however, in order to include the genetic component in the results, we need to treat separately the couples where both parents have the Neolithic marker, only one of the parents have it, or none of the parents have it (more details on this issue are given in Chapters 2 and 5).

We run our simulation for approximately 200 iterations, each corresponding to one generation. In each generation, the population reproduces, interacts culturally (i.e., some hunter-gatherers become farmers) and part of it migrates. Note that this scheme already includes the time delay between migrations, during which the children live with their parents until adulthood (when they can migrate themselves and initiate the cycle anew). This process yields an advancing front of farmers, for which we can know the relative presence of individuals with the Neolithic marker. The analysis of these data shows a clear cline on the presence of the Neolithic marker away from the region of origin. By setting the initial fraction of farmers with this Neolithic marker (haplogroup K) in the Near East in agreement with the aDNA observations and running the model for different intensities of cultural transmission, we obtain clines that can be compared to the observed one. The comparison of the modelled and observed clines enables us to find out that the importance of cultural diffusion was apparently minimal, and that only about 2% of farmers took part in cultural diffusion (this result is quantified for the first time in Chapter 5 [9]). It is important to stress that cultural transmission, albeit very weak, is still absolutely necessary to explain the existence of the observed cline. Indeed, if no hunter-gatherers became farmers, the presence of the Neolithic marker K would not diminish with increasing distance from the Near East but would be uniform. Our conclusion that the cultural effect was minimal (compared to the demic one) agrees with previous findings of genome-wide studies by other authors [80, 81, 82, 86, 87]. However, in contrast to our work (Chapter 5 [9]), genome-wide studies cannot estimate the percentage of farmers involved in cultural diffusion (because, as explained at the end of Sec. 1.1.3, the dynamics of different markers depend on different processes).

1.4. Objectives

The main goal of this thesis is to improve the understanding of several biophysical systems where reaction and diffusion processes coexist. By developing new equations and numerical simulations, it is intended to effectively and accurately explain the current experimental data in three different reaction-diffusion systems, namely virus infections (Chapter 3), virus treatments of cancer tumors (Chapter 4), and the genetics of early Neolithic populations in Europe (Chapter 5).

Each of the three systems considered requires a different mathematical model. However, reaction (interaction, reproduction) and diffusion (movement, migration) processes coexist in all of these systems. Each of the three main chapters in this thesis is devoted to a specific biophysical system, for which experimental data exist that can be directly compared to the predictions of our new models. Firstly, various strains of T7 viruses interacting with *E. coli* bacteria (Chapter 3). Secondly, Vesicular stomatitis viruses (VSVs) spreading through glioblastoma (GBM) tumor cells (Chapter 4). Thirdly, the cline of mitochondrial haplogroup K in early farming European populations (Chapter 5).

For all these systems, our main objective is achieving a better quantitative agreement between a theoretical model and experimental or observational data.

More specific objectives are the following.

- In Chapter 3, to obtain a mathematical model with full biological and mathematical meaning.

- Also in Chapter 3, to derive propagation speeds of virus infections that are consistent with the experimental data for different strains of the T7 virus infecting *E. coli* bacteria, without requiring the use of any free or adjustable parameters.

- In Chapter 4, to develop a realistic mathematical description of VSVs infecting GBMs.

- Also in Chapter 4, to explain the propagation speeds of VSVs infecting GBMs obtained in *in vitro* experiments.

- In Chapter 5, to understand the observed decrease of haplogroup K in the Neolithic populations of farmers as they migrated from Syria across Europe and interacted with Mesolithic hunter-gatherers. We also aim to understand the observed decrease of haplogroup K with the passage of time in each of the regions studied.

- Also in Chapter 5, to obtain a quantitative estimate of the intensity of cultural diffusion in the spread of the Neolithic across Europe. This intensity is defined as the average number of local hunter-gatherers who were incorporated in the farming communities per each pioneering farmer, by means of acculturation and/or interbreeding.

2. Materials and methods

This Chapter is mainly devoted to giving details on the parameters used in the next Chapters, especially how their values are calculated and what they are used for. Because in Chapters 3, 4 and 5 (which reproduce our published papers) it is not always possible to explain all parameters in detail, here we include an in-depth explanation on each of them. In addition, this Chapter also contains details on the analytical and numerical methods applied to calculate front speeds, population densities and genetic percentages in the simulations.

2.1. Data on viral infections

The two first published papers in this thesis (Chapters 3-4) deal with viral infection systems. Firstly, we consider the dynamics of a growing plaque when *E. coli* bacteria are infected by T7 viruses (Chapter 3). Secondly, the oncolytic effect of VSVs on GBMs is analyzed (Chapter 4). This section describes the main parameters related to virus infection processes, providing for each parameter a brief introduction and then details on the parameter values used in Chapters 3-4.

2.1.1. Diffusion coefficient (D)

The diffusion coefficient or diffusivity D is related to how long it takes for a particular substance (e.g., viruses, proteins, etc.) to diffuse up to a given distance through a specific medium (such as water or agar). This parameter D is encountered in all reaction-diffusion equations used in this thesis to describe the diffusion of viruses (Eqs. (1.20) and (1.23) and Chapters 3-4) and tumor cells (Eqs. (1.27)-(1.31) and Chapter 4). Therefore, it is necessary to know the value of D for the spread of viruses in the studied media (T7 in agar with *E. Coli* bacteria in Chapter 3, and VSV in a medium with GBMs in Chapter 4), as well as the value of D for the spread of GBM tumoral cells (in Chapter 4). Unfortunately, in none of these systems has the virus diffusion coefficient been directly measured. Thus, it has been necessary to use adequate proxies, as we next explain.

***D* for T7 viruses.** As previously applied in other works [4, 14, 21], because the T7 virus resembles in shape and size the P22 virus, the diffusivity for T7 virus on an agar plate should be similar to that measured for P22. Thus, we can assume that in agar $D_{T7} \approx D_{P22} = 4 \cdot 10^{-8}$ cm^2/s [4, 148]. However, in virus infections T7 viruses do not disperse freely in agar, but rather through a continuous medium of agar with a suspension of the host bacteria, *E. coli*, which adsorb the viruses. For this reason, the diffusion coefficient should be replaced by a more refined value that takes the presence of bacteria in the agar into account. We obtain the effective diffusivity under such conditions using an equation due to Fricke [149], which has been found to agree very well with experimental observations of blood cell suspensions [149, 101]. The effective diffusivity according to Fricke's equation is expressed as

$$D_{eff} = \frac{1-f}{1+\frac{f}{x}} D, \tag{2.1}$$

where f is the concentration of bacteria relative to its maximum possible value (note that if $f = 1$ then $D_{eff} = 0$), and x is related to the shape of the host cell (in this case, *E. coli*). For spherical particles

$x = 2$, which has been sometimes used in virus diffusion studies [14, 144], but for *E. coli* it is more accurate to use the value $x = 1.67$ [4, 150]. This is therefore the value of x that we use in Chapter 3 to estimate the diffusion coefficient of T7 viruses through agar with a suspension of *E. coli* cells, as a function of the relative concentration of the latter, f.

***D* for VSVs**. As mentioned above, D_{VSV} is unknown for the medium in which the propagation speed of VSV fronts infecting GBMs has been measured (this is a serum free medium [24], and details on the corresponding front speed are given in Sec. 2.1.9 below). For this reason, in Chapter 4 we use the only diffusion coefficient of VSVs that we are aware of, namely that measured in a water solution, $D_{VSV} = 8.37 \cdot 10^{-5}$ cm²/h [151]. For comparison purposes, we shall also use the value of D for virus P22 measured in agar and mentioned above [21], namely $D_{VSV} \approx D_{P22} = 4 \cdot 10^{-8}$ cm²/s $= 1.44 \cdot 10^{-4}$cm²/h [148].

***D* for GBM cells.** As mentioned above, in the plaque growth experiments that we consider (Chapter 3) bacteria are immobilized. In contrast, for VSVs infecting GBMs (Chapter 4) the GBM tumor cells can move, so their diffusion has to be taken into account (as done in previous work [45], see e.g. D_T in Eq. (1.27)). Stein et al. conducted experiments on GBM cells in a collagen gel [152]. They measured the diffusion coefficient, obtaining $D_{GBM} = 3.75 \times 10^{-6}$ cm²/h. This is the value we need for our study in Chapter 4, and has been used previously to understand the spread speed of GBM tumors [34]. In Chapter 4 we shall see that this value leads to a GBM front speed that is consistent with that measured experimentally by Stein et al. [153], again in a collagen gel. We are not aware of any measurement of D_{GBM} for the serum free medium in which the propagation speed of VSV fronts infecting GBMs has been measured [24] (Sec. 2.1.9), but there are not any experimental measurements of GBM front speeds in this medium either.

2.1.2. Rate of adsorption (k_1)

The adsorption rate, k_1 in e.g. Eqs. (1.19)-(1.21), provides a measure of the probability per unit time of adsorption of a single virus particle on a single target cell. The rate of adsorption appears both in the plaque growth problem (Sec. 1.2.3 and Chapter 3) and in the study on oncolytic treatment of cancerous tumors (Sec. 1.2.4 and Chapter 4). Below we explain how we have estimated k_1 from the results of independent experiments for the two systems we are interested in this thesis, namely T7-*E. coli* (Chapter 3) and VSV-GBM (Chapter 4).

***k_1* for T7 viruses infecting *E. coli*.** In Chapter 3 we use a value of the rate of adsorption k_1 for the T7 virus on *E. coli* bacteria previously estimated [4] from a plaque-forming assay performed in the presence of potassium cyanide (KCN), which is known to inhibit virus reproduction [154]. Therefore, as there is no creation of new viruses, the observed variation (decrease) in the concentration of free T7 phages is directly related to the adsorption on *E. coli* bacteria (see data points in Fig. 2.1, left). This simplified system with no viral reproduction can be described mathematically using, e.g., Eqs. (1.20)-(1.21) but excluding in Eq. (1.20) the term containing a spatial derivative (which vanishes because we are dealing with a homogeneous system) and the term proportional to k_2 (because virus reproduction is inhibited in this experiment). This yield

$$\frac{\partial [V](r,t)}{\partial t} = \frac{\partial [B](r,t)}{\partial t} = -k_1 [V](r,t)[B](r,t). \tag{2.2}$$

From Eq. (2.2), it follows that viruses and bacteria are related by $[B] = [V] + \xi$, where ξ is a constant. For the experiment we are interested in, the value of ξ can be obtained from the initial concentrations of viruses and bacteria, yielding the value $\xi = 1.39 \times 10^8$ ml^{-1} [4]. Moreover, integrating Eq. (2.2) for the concentration of viruses,

$$\ln\left(\frac{[V]+\xi}{[V]}\right) - \ln\left(\frac{[V]_0+\xi}{[V]_0}\right) = k_1\xi(t-t_0), \tag{2.3}$$

where $[V]_0$ is the value of $[V]$ at $t = t_0$. Fitting Eq. (2.3) to the experimental data in Fig. 5 in Ref. [154] (shown in Fig. 2.1, left), the slope yields $k_1 = (1.29 \pm 0.59) \times 10^{-9}$ ml/min [4].

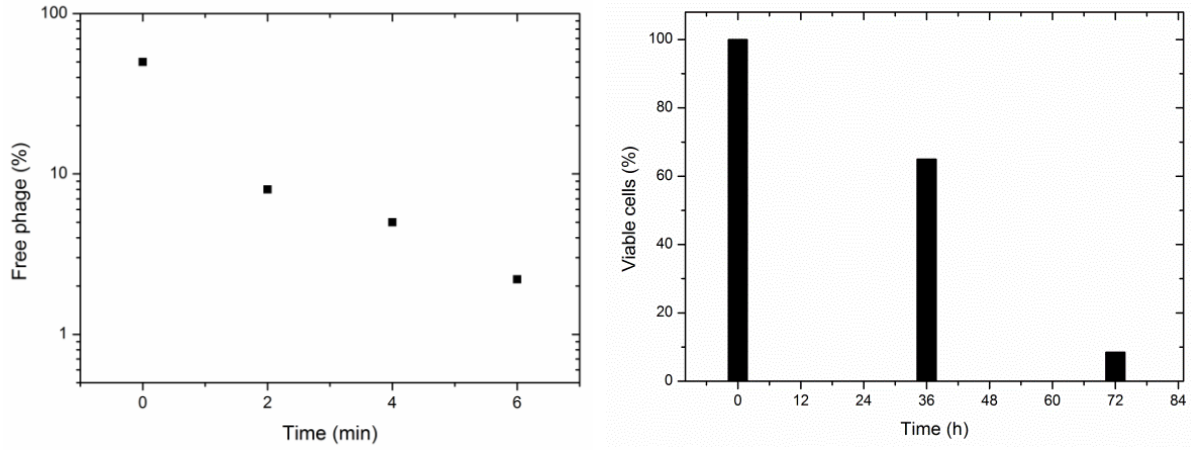


Figure 2.1 Left: T7 adsorption on *E. coli* (at $\boldsymbol{t = -2}$ min the percentage of free phage was 100%, from $\boldsymbol{t = -2}$ min until $\boldsymbol{t = 0}$ the sysem was kept on ice, and then transferred at 37ºC). Adapted from Fig. 5 in Ref. [154]. Right: Cell viability of human GBM cells after 36 and 72 hours post-infection with VSV (variant G/GFP). Adapted from Fig. 3C in Ref. [25].

k_1 for VSVs infecting GBMs. To estimate the value of the rate of adsorption of VSVs on GBMs (which will be used in Chapter 4), there is unfortunately no experiment where the reproduction of VSVs is inhibited (such an experiment would be analogous to that for T7-*E. coli* described in the previous paragraph). Thus we have used data from a cell viability experiment of GBMs after infection with VSVs (Fig. 2.1, right), where virus reproduction was not inhibited. To minimize the possible effect of virus reproduction, we used only the earliest recorded measurement, namely at 36 hours post-inoculation (Fig. 2.1, right). In this way, it seems reasonable to neglect virus reproduction, because in the next section we shall find that virus reproduction takes place after about 2 days, which is longer than 36 hours. Thus, similarly to the experiment for the T7-*E. coli* system (previous paragraph), we can assume that the dynamics follows Eq. (2.2) and therefore $[T] = [V] + \xi$ (we denote the tumoral cells as T rather than B). Because for this experiment we have data on the concentration of tumoral cells (rather than viruses), we use $[V] = [T] - \xi$ into Eq. (2.3) and obtain for k_1

$$k_1 = \frac{1}{\xi(t-t_0)}\left[\ln\left(\frac{[T]}{[T]-\xi}\right) - \ln\left(\frac{[T]_0}{[T]_0-\xi}\right)\right]. \tag{2.4}$$

The value of the constant ξ can be estimated directly from the fact that in this experiment [25] the multiplicity of infection is 0.5, i.e. $[V]_0 = 0.5[T]_0$ and therefore $\xi = [T]_0 - [V]_0 = 0.5[T]_0$. The initial concentration of tumor cells is difficult to estimate because it is not directly reported by Wollmann et

al. in Ref. [25]. However, after extrapolating values for similar experiments by the same authors [24, 155], we believe that a realistic range is $[T]_0 = 10^6 - 10^8$ cells/cm^3. From the data in Fig. 2.1 (right) we see that the concentration of tumor cells at $t - t_0 = 36$ h is $[T] = 0.65\ [T]_0$. As mentioned above, we do not use the data at 72h because subsequent infection (after virus reproduction) may be playing a role. Using these values into Eq. (2.4), we estimate that the rate of adsorption of a VSV-GBM plaque lies in the range $10^{-10} < k_1 < 10^{-8}$ ml/h. Although this is a wide range, in Chapter 4 we shall find that this parameter (k_1) has, in fact, little effect on the predicted speed of the infection front.

The data shown in the right of Fig. 2.1 have been obtained for the same VSV variant (G/GFP) as that for which the front speed of VSVs infecting GBMs can be estimated (see Sec. 2.1.9 below). The advantage of this variant is that GBM cells infected by it express a fluorescent protein, which makes it easy to track infections (this will be explained in Sec. 2.1.9 and Fig. 2.3).

2.1.3. **Rate of death of infected cells (k_2)**

The knowledge of the death rate of infected cells k_2 is crucial to properly understand viral dynamics, and therefore this parameter appears in all systems in this thesis dealing with viral infections (e.g., Eqs. (1.19)-(1.28)). Because for each infected cell that dies, an average number of Y viruses are ejected (where Y is called the yield or burst size, see Sec. 2.1.5 for more details on this parameter), the rate of death of infected cells is also indicative of the reproductive rate of the virus population [156, 4, 21, 144]. For this reason, in all the models in this thesis, k_2 is encountered in both the terms related to infected cell death (e.g., in Eq. (1.19)) and those related to virus reproduction (e.g., in Eq. (1.20)). The values for the rate of death k_2 of *E. coli* bacteria infected by T7 viruses (Chapter 3) and for GBM cells infected by VSVs (Chapter 4) have been obtained from experimental data on the concentration of viruses versus time, as we explain in the following paragraphs.

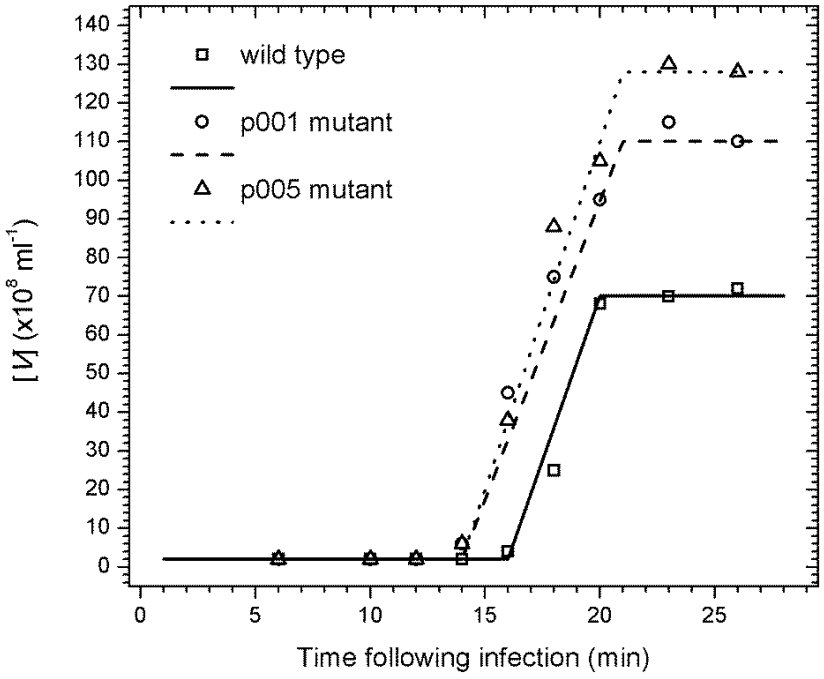


Figure 2.2 One-step growth of three T7 virus strains on *E. coli* host bacteria versus time. According to the author of Ref. [15], the concentration of host cells is uniform and adsorption is negligible, $\boldsymbol{k_1 \approx 0}$. This figure is the same as Fig. 3.1.

k_2 for T7 viruses infecting *E. coli*. We will need this parameter in Chapter 3. The death rate of viruses can be easily estimated from the one-step growth experiment, which we already mentioned in Sec. 1.1.1. This experiment is set so that initially there are only infected cells in the system. Therefore, any increase in the virus concentration will be due to the death of infected cells. This experiment always shows the same behavior, namely the concentration of viruses (initially, those infecting the cells) remains constant for a period τ (the time delay or eclipse period during which viruses reproduce within the cells, see Secs. 1.1.1 and 2.1.8), after which cell death and virus release begins, so that the concentration of viruses increases rapidly (following a step-like growth process), and finally a plateau is reached once all cells have lysed [10] (see Fig. 2.2). The rate of death of infected cells k_2 can be obtained from the virus growth interval as follows.

In the one-step experiment, if the initial concentration of infected cells is I_0, for $t \geq \tau$ (i.e., during the virus growth interval) the concentration of infected cells decays according to $d[I] = -k_2[I]_0 dt$ (see, e.g., Eq. (1.19) for the case where no new cells become infected, $k_1 = 0$). Since for each infected cell, *Y* new viruses are released, we also have that $d[V] = -Yd[I] = k_2 Y[I]_0 dt = k_2 V_{max} dt$. Therefore, during the virus reproductive interval ($t \geq \tau$), the concentration of viruses can be described by

$$[V] = k_2[V]_{max} \cdot (t - \tau) + [V]_0, \tag{2.5}$$

where $[V]_0$ is the initial concentration of viruses (i.e., for $0 \leq t \leq \tau$). If we define t^* as the time when the plateau is reached (i.e., the moment when all infected cells have lysed), from Eq. (2.5) follows that k_2 can be obtained from

$$k_2 = \frac{[V]_{max} - [V]_0}{[V]_{max} \cdot (t^* - \tau)}. \tag{2.6}$$

In Chapter 3 we need to know the death rate of *E. coli* bacteria when infected by T7 viruses. Yin [15] performed the one-step growth experiment for three strains of the T7 virus on *E. coli*, and the corresponding data are shown in Fig. 2.2. Using the data from Fig. 2.2 into Eq. (2.6), we estimate the values of k_2 for each strain as $k_2(wild) \approx 0.25$ min^{-1} and $k_2(p001) \approx k_2(p005) \approx 0.17$ min^{-1}.

k_2 for VSVs infecting GBMs. In Chapter 4 we will need numerical values for the rate k_2 of death of tumor cells of GBM infected with VSVs. Wollman et al. [25] performed an experiment to visualize the self-amplification of different VSV variants in a medium of GBM cells versus time (Fig. 4 in Ref [25]). Because the authors were not interested in performing a one-step growth experiment *per se*, their data are not as detailed as those available for the T7 strains, and they reported only three measurements at 1, 2 and 3 days post infection (dpi). Nonetheless, we can estimate the value of k_2 from those experiments. We are interested in the G/GFP variant of VSV because is that for which the front speed can be estimated (Sec. 2.1.9 below), but the data from the replication-restricted variants (labelled dG-GFP and dG-RFP) allow us to estimate the initial concentration as $10 < [V]_0 < 100$ PFU/ml (Ref. [25], Fig. 4). Here PFU stands for plaque-forming units, which is the same as viruses. On the other hand, the data for the variant G/GFP presents a maximum concentration of $10^8 < [V]_{max} < 10^9$ PFU/ml at 2 days post infection (dpi) (after which there is a decay, probably due to the start of a new infection cycle). Because of the low number of measurements, we cannot be completely sure that the maximum is reached at 2 dpi, and neither that the observed concentrations correspond to its

absolute maximum. However, the data show clearly that $[V]_0 \ll [V]_{max}$, which allows us to simplify Eq. (2.6) as:

$$k_2 \approx \frac{1}{t^* - \tau}. \tag{2.7}$$

From the fact that the observed maximum is reached at about 2 dpi, we can assume that $t^* = 48 \pm 12$ h. The value of the delay time τ (time elapsed between viral adsorption and the release of the first progeny phage) is not well defined, but we can estimate a wide range $2 < \tau < 12$ h (see Sec. 2.1.8 below). Therefore, in Chapter 4 we will consider k_2 as a function of τ. Note that the aforementioned ranges of t^* and τ imply that $0.017 < k_2 < 0.042\ \mathrm{h}^{-1}$.

2.1.4. Rate of death of viruses (k_3)

In a viral infection system, the concentration of free viruses can decrease not only as a result of the infection process, but also because of natural death, which is characterized by the death rate of viruses k_3.

k_3 of T7. In the models describing the plaque growth problem of virus T7, parameter k_3 is never included (see, e.g., Eqs. (1.19)-(1.26)) [14, 4, 21], because the death rate is substantially lower than the growth rate. For example, the growth rate of the wild T7 viruses considered in Chapter 3 would be $k_2 Y = 0.25 \cdot 34.5 = 8.63$ min^{-1} (see Secs. 2.1.3 and 2.1.5), which is much higher than the death rate reported by Arnold et al. [157] for a T7 virus, namely $k_3 = 0.014$ min^{-1}. Accordingly, in Chapter 3 we will assume that $k_3 \approx 0$.

k_3 of VSV. Also in the study of oncolytic treatments, the virus death rate is sometimes low. However, previous authors have included the effect of k_3 in their models, see Eq. (1.29). Similarly, in Chapter 4 we also include the natural death of VSVs. To estimate the value of k_3, we need data on the evolution of the concentration over time from an experiment in which viruses cannot replicate. Such data are available from Fig. 4 in Ref. [25], which refer to two replication-restricted VSV variants (i.e., the same plots that we have used in Sec. 2.1.3 to estimate $[V]_0$). Under these conditions, the equation which rules the dynamics of the virus population is

$$dV = -k_3 V dt, \tag{2.8}$$

which after integration yields

$$V(t) = V_0 e^{-k_3 (t - t_0)}. \tag{2.9}$$

In agreement with Eq. (2.9), we can estimate the value of k_3 from the slope of the linear regression of $\ln V$ versus t. In this way, using the data mentioned above, we obtain the range $0.014 < k_3 < 0.028\ \mathrm{h}^{-1}$.

2.1.5. Burst size of viruses (Y)

When a virus adsorbs to a host cell, it infuses the cell with its nucleic acid and replicates within. Eventually, the new progeny leaves the cell. The number of viruses ejected from a single cell is called the yield or burst size, Y. Therefore, since $k_2[I]$ is the number of cells that die per unit time (see Eq.

(1.19)), the number of viruses appearing per unit time is given by $k_2 Y[I]$ (see Eq. (1.20)). Thus, knowing Y is very important to describe the virus infection dynamics in the problems studies in Chapters 3 and 4. As we detail below, the burst size can be estimated from the same experiments used in Sec. 2.1.3 to estimate the death rate k_2 of infected cells.

Y* for T7 infecting *E. coli. Numerical values of this parameter will be applied in Chapter 3. The value of Y can be easily estimated from the one-step growth experiment described in Sec. 2.1.3, thus we use the same experimental data in Fig. 2.2 (obtained from Ref. [15]) to estimate the burst size of T7 replicating on *E. coli*. At the end of the one-step growth experiment, the concentration of viruses V_{max} corresponds to the global yield of the initially infected cells I_0, all of which have lysed. Therefore, from the definition of Y above it follows that $Y = \frac{V_{max}}{I_0}$. At the beginning of the experiment there are no free viruses, and the experimental points in Fig. 2.2 before the start of the growth process correspond to the concentration of infected cells [10, 4]. Thus we can estimate the yield or burst size from the data in Fig. 2.2 as $Y = \frac{V_{max}}{V_0}$. In this way, we obtain for the three T7 strains in Fig. 2.2 that $Y(wild) = 34.5$ [4], $Y(p001) = 56.5$ and $Y(p005) = 65$.

***Y* for VSVs infecting GBMs**. We will need this parameter in Chapter 4. As mentioned in Sec. 2.1.3, there is no available data on any one-step growth experiment performed for VSVs infecting GBM cells. However, as in Sec. 2.1.3, we can use as a proxy the data on VSV replication in GBM cells published in Ref. [25], Fig. 4 therein. Several VSV variants were reported in Ref. [25]. However, recall that we are interested in the G/GFP variant because the front speed can be estimated for this one (this is done in Sec. 2.1.9 below). As already explained in Sec. 2.1.3, we have estimated the initial concentration to be in the range $10 < [V]_0 < 100$ PFU/ml, and the maximum concentration of viruses (after all initially infected cells have lysed) $10^8 < [V]_{max} < 10^9$ PFU/ml. Assuming that all viruses infect a cell, and applying that $Y = \frac{V_{max}}{V_0}$, we obtain the range $10^6 < Y < 10^8$ for the burst size of VSVs replicating in GBM tumor cells. This range is similar to that measured for VSVs infecting hamster kidney cells [158] (Ref. [159] also includes some one-step curves for that system, but those in Ref. [158] are substantially more precise because they contain more data points).

2.1.6. Proliferation rate of cells (a)

Tumor cells reproduce very fast [160]. Their dynamics usually follows logistic growth [45, 34], Eq. (1.8), characterized by the proliferation rate a (see, e.g., Eq. (1.27)). This process refers to the net reproduction of tumor cells, i.e., it includes the effects of their fast multiplication and also the natural death of cells or apoptosis. The proliferation rate a is necessary in our models in Chapter 4 to estimate the predicted speed of tumor growth.

***a* of GBM tumors**. The proliferation rate a of GBM tumor cells has been measured *in vitro* [153] and *in vivo* [161]. We next summarize both kinds of data.

Based on *in vitro* data, cell doubling times can be estimated, and Stein et al. used them to suggest the range $0.04 < a < 0.3$ day^{-1} [153]. They also conducted experiments with GBM spheroids in a collagen gel supplemented with a minimum essential medium and fetal bovine serum (the latter is necessary for cells to reproduce, whereas the former is not and is thus called minimum essential, or serum free medium). In this way, Stein et al. could compare observed tumor front rates and profiles

to those predicted by a reaction-diffusion model of tumor growth. This comparison confirmed the *in vitro* range $0.04 < a < 0.3$ day^{-1} [153].

It could be argued that, if an experiment were performed *in vitro* so that we could simultaneously observe the virus front and the tumor cell front (as depicted in Fig. 4.1), the presence of growth supplement (fetal bovine serum) might perhaps lead to a different value of D_{VSV} than those reported in Sec. 2.1.1 above, and also to a different front speed than that reported in Sec. 2.1.9 below (because the latter measurements were made in serum free medium [24]). However, concerning the front speed (Sec. 2.1.9), the difference is likely to be small because the virus front propagates in a region with tumor cells at carrying capacity, i.e. in which most (if not all) of the growth medium has been already consumed by tumor cells. For the same reason, the effect of the growth serum on the value of D_{VSV} (Sec. 2.1.1) is likely to be negligible (recall also that in Sec. 2.1.1 we have had to approximate the value of D_{VSV} to its values in the few media in which it has been directly measured).

Concerning *in vivo* data, Rockne et al. [161] applied a simple biologically-based reaction-diffusion model to describe the observed growth of gliomas in human patients. The proliferation rate was estimated by fitting the model to actual *in vivo* data from nine patients diagnosed with GBM, yielding a range of *in vivo* values $0.01 < a < 0.14$ day^{-1} [161].

Combining both results, in Chapter 4 we use the range $0.01 < a < 0.3$ day^{-1}, and assume that $a = 0.1$ day^{-1} is a reasonable mean value.

2.1.7. Saturation cell density (k)

The saturation cell density k is the maximum concentration of cells attainable in a culture vessel, under specified conditions, i.e. the maximum number of cells per unit of volume that a system can support (this value is equivalent to the concept of carrying capacity in ecological dynamics). The saturation cell density k is necessary to set a limiting bound for the cell growth when it follows logistic dynamics, as usual for growing tumors [45, 34]. Therefore, parameter k appears in the models describing oncolytic dynamics, e.g. in Eq. (1.27). Since the models in Chapter 4 describe the oncolytic effect of VSVs on a growing GBM tumor, we will need to know the value of k for GBM cells.

In contrast, in Chapter 3 we consider T7 viruses infecting *E. coli* bacteria which have depleted their nutrient and, therefore, cannot grow in number [12]. Thus, for the T7-*E. coli* system (Chapter 3) there is no cell reproduction, and we do not need to estimate the saturation density k (this is why the logistic term does not appear in, e.g., Eq. (1.21)). Remember that (as explained in Sec. 1.1.1), unlike other viruses, T7 infections do not cease when bacteria exhaust their nutrient, which facilitates the observation plaque growth and front speed measurements [12].

***k* of GBM cells.** This information was not measured in the specific oncolytic experiments to which we apply our models [25, 24, 155]. Therefore, in Chapter 4 we use the value $k = 10^6$ cells/cm^3, which has been previously measured [162] for GBM cells and used in reaction-diffusion modelling studies [163, 162].

2.1.8. **Delay time (τ)**

When mathematically modelling virus infections, the delay time τ explicitly accounts for the time interval during which viruses are replicating inside the cell, i.e., the time elapsed between the moment a virus adsorbs on a cell and the moment it releases the new progeny. During this interval, neither the original virus nor its progeny diffuses, so τ has an important effect on the reaction-diffusion dynamics of an infection front, as discussed in Sec. 1.2.2. The first models developed to describe growing plaque dynamics did not include this effect (see Eqs. (1.19)-(1.21)). Therefore, we will base our models in Chapters 3 and 4 on the time-delayed model by Amor and Fort [21], i.e. Eqs. (1.22)-(1.26), which include the effect of τ explicitly.

τ* of T7 infecting *E. coli. We will need the value of the delay time τ for this system in Chapter 3. We can estimate it by resorting again to the one-step experiment in Fig. 2.2 [4], which has been already used above to estimate the values of k_2 and Y. In this experiment, the delay time τ can be measured directly from the data as the time elapsed between adsorption ($t = 0$) and the first release of viruses (i.e., the beginning of the rise in the virus concentration). The values of the delay time used in Chapter 3 for the three different strains in Fig. 2.2 thus obtained are $\tau(wild) = 16$ min and $\tau(p001) = \tau(p005) = 14$ min.

***τ* of VSV infecting GBM.** This parameter will be used in Chapter 4. Estimating the time delay for the VSV infecting GBM tumor cells is not as straightforward as for T7 infecting *E. coli*, because no data on one-step growth experiments is available. However, we can estimate upper and lower bounds for τ from two results obtained by Wollman et al. [24] when studying the oncolytic potential of several viruses, including VSV. Firstly, the death of infected cells starts about 6 hours after the inoculation of viruses, but it takes 4 h since the inoculation and the first evidence of infection (which is visible because they use fluorescence altered virus strains). Therefore, $\tau > 2$ h. Secondly, in another experiment infected cells are added directly (rather than free viruses) and newly infected cells appear about 12 h later. Thus $\tau < 12$ h. Combining both results we obtain a wide range for the delay time in the VSV-GBM system, $2 < \tau < 12$ h. For this reason, the results in Chapter 4 will be presented as a function of τ (Fig. 4.2).

2.1.9. **Front speeds of viral infections (c)**

After inoculating viruses in a small area of a medium containing susceptible cells, an infection front is formed which expands outwards with a constant front speed, which can be measured experimentally. In this thesis we aim to develop realistic reaction-diffusion models that can estimate the observed speeds of the infection fronts of T7 viruses on *E. coli* (Chapter 3) and of VSV on GBM tumoral formations (Chapter 4). Therefore, we need to have experimental measurements of the front speed that we want to explain.

***c* of T7 fronts on *E. coli*.** Bacteriophages are viruses that infect bacterial cells. In Chapter 3 we will analyze bacteriophage T7 infecting *E. coli* bacteria. Yin [15] took screenshots of growing plaques of Y7 infecting *E. coli* at different times (Fig. 1 in Ref. [15]). As time passes, the plaque diameter increases and, by measuring the diameter at different times (13, 18 and 23h post-inoculation), Yin could estimate average speed of the infection front [15] for three different strains of T7 (namely wild, p001 and p005, i.e. the same strains in Fig. 2.2). In this way, he obtained $c_{wild} = 0.195 \pm 0.012$ mm/h,

$c_{p001} = 0.253 \pm 0.009$ mm/h and $c_{p005} = 0.249 \pm 0.009$ mm/h. These are the speeds that we aim to explain with the models developed in Chapter 3.

***c* of VSV fronts infecting GBM tumors.** VSVs infect mammalian cells, not bacteria. The oncolytic properties of VSVs against tumors have been studied in many scientific contributions, e.g. in Refs. [159, 25, 24, 26, 164, 23]. We will contribute to this topic in Chapter 4. However, in contrast to the T7-*E. coli* system, the speed of VSVs infecting a GBM tumor has not yet been measured using experiments specifically designed to this end. However, Wollman et al. [24] developed an experiment to assess viral replication and spread of VSVs (among other oncolytic viruses) targeting GBM cells, from which we can estimate the speed of the infection front. This experiment is essentially a transfer of infected cells grown on a glass chip to a non-infected culture of GBM cells in serum-free medium. Then, it is possible to track the spread of the infection front because viruses have been altered so that the infected cells emit fluorescence. A screenshot of this experiment was taken at 24 hours post-infection, which is shown in Fig. 2.3 (adapted from Fig. 3A in Ref. [24]). From this image we know that at $t = 0$ the infection front is delimited by the glass chip (red lines in Fig. 2.3), and that the infection has spread at $t = 24$ h approximately up to the green circle highlighted in Fig. 2.3. From the distances between the position of the front at these two different instants, we can estimate a range of experimental values for the front velocity. In this way, we obtain that $c_{VSV} = (4 - 5.4) \cdot 10^{-3}$ cm/h, which is the range of speeds we aim to explain with our models in Chapter 4.

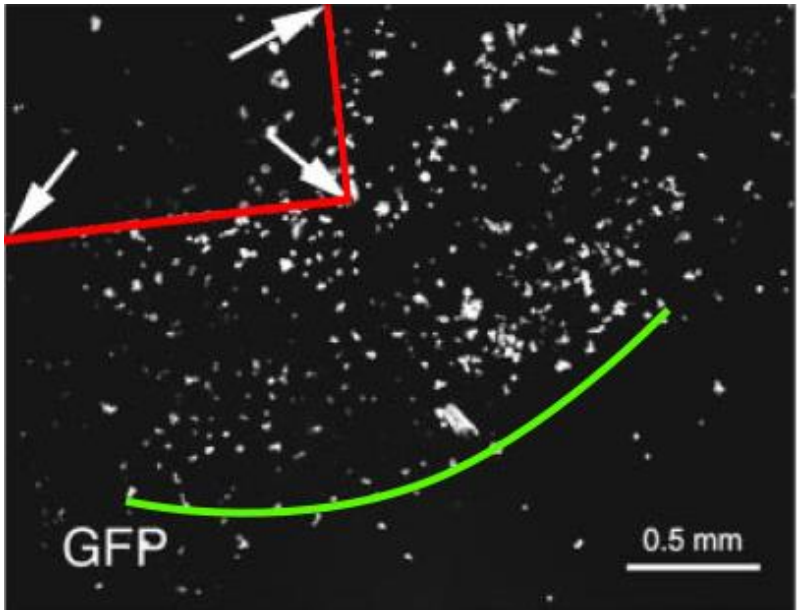


Figure 2.3 A fluorescent protein expressed by GBM cells infected by the variant of VSV called G/GFP makes it possible to track infections in a culture dish using a fluorescence microscope. White arrows and red lines show the location of the small glass chip carrying the initially infected cells, i.e. the origin of the infection. The green curve shows the average displacement for viruses 24 h after the start of the infection. Original image extracted from Fig. 3A in Ref. [24], on which we have added the red and green lines.

2.2. Data on pre-industrial human populations

2.2.1. Ancient DNA data

Ancient DNA (aDNA) is DNA isolated from ancient biological material. A genetic cline is a gradual spatial variation in the frequency of a genetic marker. Chapter 5 notes the existence, and seeks an explanation, for a human aDNA cline in Neolithic populations through Syria, Anatolia (present-day Turkey) and Europe. To this purpose, in Chapter 5 we develop demographic and genetic simulations in space and time. We have detected the cline using a genetic database that we have gathered from various aDNA studies, e.g. Refs. [78, 165, 166, 76, 83, 167, 168]. In those works, samples were recollected from skeletal human archaeological material (usually bone and teeth). Chapter 5 performs the simulations at the continental scale, because aDNA data are not yet numerous enough to study local geographic regions. When geneticists analyze aDNA from human remains, they often report the mitochondrial DNA (mtDNA), which is inherited through the maternal route and is found in both males and females. Fewer studies deal with the Y chromosome (which is inherited paternally and found only in male individuals). Chapter 5 considers mtDNA because there are more data for it, and this will make it possible to attain better statistics (i.e., narrower error bars). In the future, we hope that our approach will be applied not only to other mtDNA clines, but also to Y-chromosome clines.

As explained in Sec. 1.1.3, an haplogroup is a group of several DNA sequences (haplotypes) with a common ancestor. With regard to Chapter 5, it is important to mention that the mtDNA of European hunter-gatherers is composed mainly of U lineages (haplogroups U, U4, U5, and U8), which are absent in the early Neolithic (i.e., farmer) populations in Europe. In contrast, haplogroups N1a, T2, K, J, HV, V, W, and X have been found in early Neolithic European farmers [78, 169]. This set of mtDNA haplogroups (together with some Y-chromosome haplogroups) is sometimes called the Neolithic 'genetic package' [78], in analogy to the Neolithic 'archaeological package' (i.e., cereals, sheep, ceramics, etc.) [170].

An exhaustive search has been necessary to compile the database used in Chapter 5 (which is included as Appendix A to this thesis), because there was so far no such an exhaustive database available, and new data were being continuously published. In this way, we have been able to gatherer the information of 513 Neolithic individuals from remarkable scientific reports (mtDNA haplogroup, date, latitude, longitude…) and grouped them into 26 regional cultures according to their geographical and cultural closeness (plus five individuals which could not be grouped in regional cultures of more than two individuals). Figure 2.4 shows the 26 regions (used in Chapter 5) on a map of Europe and the Middle East.

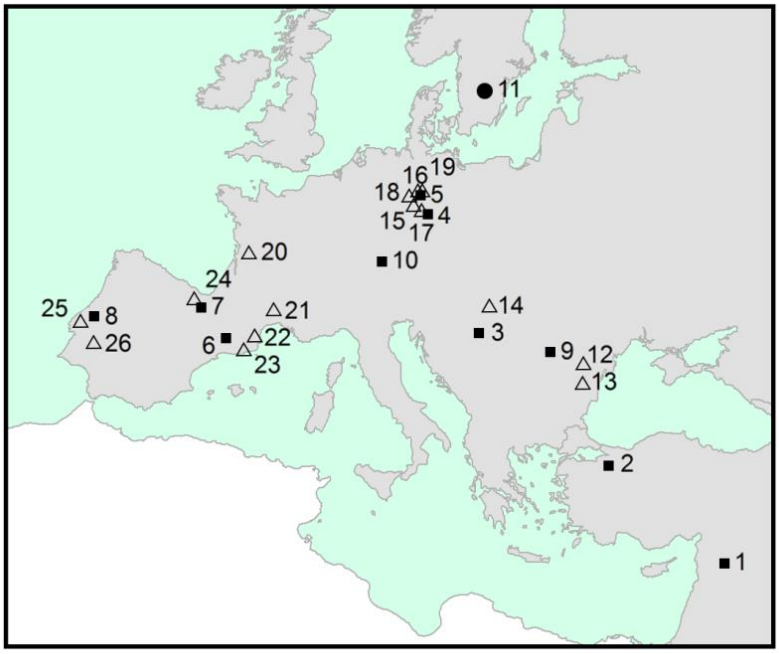


Figure 2.4 Location of the 26 regional cultures with ancient mtDNA data used in the study of Chapter 5. Squares correspond to the oldest regional Neolithic cultures (first arrival wave of farmer settlers to that area), namely 1 Syria PPNB, 2 Anatolia, 3 Hungary-Croatia Starčevo, 4 Eastern Germany LBK, 5 Western Germany LBK, 6 Northeastern Spain Cardial, 7 Spain Navarre, 8 Portugal coastal Early Neolithic, 9 Romania Starčevo, 10 Southern Germany LBK, and 11 Sweden (circle). Triangles correspond to more recent regional cultures, namely 12 Romania Middle Neolithic, 13 Romania Late-Middle Neolithic, 14 Hungary LBK, 15 Eastern Germany RSC, 16 Eastern Germany SCG/BAC, 17 Eastern Germany SMC, 18 Western Germany BAC, 19 Western Germany BEC, 20 Western France Prissé, 21 South-Eastern France Treilles, 22 Catalonia Epicardial, 23 Catalonia Late Epicardial, 24 Spain Basque country, 25 Portugal coastal Late Neolithic and 26 Portugal inland Late Neolithic.

For each of the 26 regions in Fig. 2.4, the percentage of haplogroup K is calculated by dividing the number of individuals with haplogroup K by the total number of individuals whose mtDNA haplogroup is known. For the individuals of each of the 26 regions, we also find their average date, as well as their average great-circle distance to the site of Ras Shamra, Syria (this is the oldest Neolithic Syrian site in previous work [47], and we therefore consider it as a reasonable origin for the Neolithic wave of advance). Of these 26 regions, we select 9 for our quantitative study. Firstly, because they are the oldest Neolithic regional cultures for which there are genetic data available, so they are the best ones to analyze the result of the Neolithic spread before other processes could affect the genetic pattern (e.g., further interbreeding between farmers and hunter-gatherers, population movements, etc.). And secondly, because each of these regions includes at least 8 individuals (regions with fewer individuals are ignored to avoid large error bars). These 9 regions are 1. Syria PPNB [165], 2. Anatolia (current Turkey) [80, 171], 3. Hungary and Croatia Starčevo [172, 83], 4 and 5: Germany LBK (Eastern and Western) [78, 80, 169, 83, 76], 6. North-Eastern Spain Cardial [83, 166, 173], 7. Navarre (Northern Spain) [78, 167], 8. Portugal [78, 173, 174, 175] and 11. Sweden TRB [176, 177]. A more detailed discussion will be given in Chapter 5 (Secs. 5.8.2 and 5.8.10).

As mentioned above, the database used in Chapter 5 is included as Appendix A Data S1 to this thesis. Table 2.1 shows, as an example, one individual and some of its important features for each of the 9 regions used in the quantitative study in Chapter 5.

iD	Location	Lat., Long.	BCE y	mtDNA	Archaeol. context	Ref.
H25	Tell Halula (Syria)	36.416, 38.166	7400	K	Middle PPNB	[165]
BAR26	Bartin (Turkey)	40.3, 29.5666	6300	N1a1a1	Anatolia Neo	[80]
BAM14	Alsónyék-Bátaszék, (Hungary)	46.205, 18.705	5685	J1c	Starčevo	[172]
KAR7	Karsdorf (Germany)	51.273, 11.656	5137.5	K1a	LBK	[78]
HAL21	Halberstadt-Sonntagsfeld (Germany)	51.89, 11.04	5137.5	T2b	LBK	[78]
CSA26	Can Sadurni, Barcelona (Spain)	41.334, 1.922	5390	X1	Cardial culture	[166]
CAS204	Los Cascajos, Navarra (Spain)	42.559, -2.188	4932.5	U5	Early Iberian Neo	[167]
F19	Almonda cave (Portugal)	39.505, -8.615	5265	H4a1a	Early Neolithic	[173]
Res15	Resmo (Sweden)	56.538, 16.446	2651	J1d5	TRB	[176]

Table 2.1 Some individuals of the Neolithic mtDNA database gathered for the study in Chapter 5 (the complete database is included as Appendix A to this thesis).

2.2.2. Persistence (p_e)

Let the persistence p_e of a population stand for the proportion of individuals that reproduce at the same location where they were born. Then, $(1-p_e)$ indicates the probability to move from the birthplace to other places. The persistence p_e is an important parameter in some population dispersal models, e.g. Eq. (1.33). An ethnographic study of the Majangir people, which are pre-industrial agriculturalists in Ethiopia, provides the persistence of 3 groups, namely 0.54, 0.40 and 0.19 [178]. As in previous work [139, 179, 94, 47], in Chapter 5 we will use the average of these 3 values, namely $p_e = 0.38,$ as a representative value to model the spread of the Neolithic. The main difference to those earlier works is that, besides the population density, we will also model the genetics.

2.2.3. Mobility (m)

Let the mobility m of a population stand for the mean squared displacement per generation, i.e.

$$m = \frac{\langle \Delta^2 \rangle}{T}, \tag{2.10}$$

where Δ^2 is the square of the distance between the birthplace of a parent and that of one of her/his children, the mean squared displacement $\langle \Delta^2 \rangle$ is the average of Δ^2 over all individuals of the population considered, and T is the generation time (defined as the age difference between the parent and her/his child, see Sec. 2.2.5). The interest of the mobility in Neolithic spread models is that the first such mathematical models were based on diffusion equations, e.g. on the Fisher equation (1.9). In that framework, we know from Secs. 1.2.1 and 1.2.2 that, in two-dimensional space, the

diffusion coefficient D is simply related to the mobility m as $D = \frac{m}{4}$ (note that here we use the symbol T rather than τ, which has been used in Secs. 1.2.1.2, because T is the usual notation for human populations [139, 179, 94, 47, 9], whereas τ is used in virus studies [4, 115, 21, 7, 8]). Ammerman and Cavalli-Sforza [52] estimated the mean squared displacement per generation for the three groups of pre-industrial farmers mentioned above [178], obtaining $m_1 = 1115.7$ km^2/gen, $m_2 = 1325.6$ km^2/gen and $m_3 = 2153.0$ km^2/gen. Fort and Méndez [51] made an statistical analysis of a set of values (including those three) and obtained $m = 1544 \pm 368$ km^2/generation (80% confidence-level interval). This range has been also used in further Neolithic transition studies [139, 179, 94, 47]. Similarly, here we will also use the average value $m = 1544$ km^2/generation (or $\langle \Delta^2 \rangle = 1544$ km^2).

As already mentioned in Sec. 1.3.3, the computational model in Chapter 5 uses a grid of square cells to subdivide the map of Europe and Near East. In that model, the dispersion of farmers in each iteration (generation) of the simulation takes place towards the four nearest neighboring cells (i.e., to the nearest ones to the north, south, east and west) of the original cell. This simple computational model makes simulations substantially faster, and we expect that more complicated dispersal models (with several distances and probabilities) would not change our conclusions. Then, since a fraction p_e of the population does not move and the remaining fraction, $(1 - p_e)$, moves a distance r (which is the length of a side of the square cells), we can obtain a reasonable estimate of r by requiring that the mean squared distance moved per generation is

$$(1 - p_e) r^2 = \langle \Delta^2 \rangle, \tag{2.11}$$

where, as seen above, $p_e = 0.38$ is the persistence (Sec. 2.2.2) and $\langle \Delta^2 \rangle = 1544$ km^2. From Eq. (2.11), we obtain that $r \approx 50$ km. This is the length of the sides of the square cells in our computational model in Chapter 5. Thus, our simulation grid is made of cells with area $50 \times 50 = 2500$ km^2. Interestingly, a previous genetic simulation study (which did not perform the detailed estimation above for r) also used a square grid of cells with side $r = 50$ km [89].

2.2.4. Maximum population densities of farmers ($p_{F\,max}$) and hunter-gatherers ($p_{HG\,max}$)

In the simulations we need to limit the number of individuals present in a cell after population dispersion or growth [see Eq. (1.32)]. For this reason, we need estimates for the maximum population densities of farmers ($p_{F\,max}$) and hunter-gatherers ($p_{HG\,max}$), which are also called carrying capacities or saturation densities. The transition from hunting and gathering to farming and stockbreeding led to significant increases in the population density [52, 49]. Ethnographical estimations of population densities for hunter-gatherers (HG) vary widely (for a table of ranges in different habitats see, e.g., Ref. [180]). Based on such data [180, 181], a representative value that has been used in previous simulations (already cited in the previous paragraph) of the spread of the Neolithic across Europe [89] is $p_{HG\,max} = 0.064$ HGs/km^2. The population density of pre-industrial farmers is substantially higher. For example, the value that was used in those previous simulations [89] is 20 times that for hunter-gatherers, i.e. $p_{F\,max} = 1.28$ farmers/km^2. We shall apply these values as reasonable estimates of the saturation density, which we shall use to compute the maximum cell population. As explained in Sec. 2.2.3, each cell of the grid in the computational model (Chapter 5) covers an area of 50×50 km^2, which means that cells that can be inhabited (i.e., those not located on the sea) have a maximum population number of $P_{F\,max} = 3{,}200$ farmers/cell and $P_{HG\,max} = 160$ HGs/cell (for farmer and hunter-gatherer populations, respectively).

2.2.5. **Generation time (T)**

In simulations of human range expansions, the generation time T is defined as the mean time interval between the migration of an individual and one of her/his children. Note that we do not consider the oldest child but the average over all of them, because this is the relevant quantity driving the dynamics of front propagation (for a detailed analysis, see Ref. [182]). In the absence of ethnographic data that allow a direct estimation of this time interval for pre-industrial populations, the mean age difference between an individual and her/his child is used instead [182, 47]. In the computational model used in Chapter 5, each iteration includes the dispersal, cultural interaction and reproduction of individuals belonging to the same generation. Therefore, the time interval between two successive interactions is equal to the generation time T. Ethnographic observations of preindustrial farmer populations [178] lead to the probabilities $p_1 = 0.46$ for $T_1 = 27$ yr, $p_2 = 0.51$ for $T_2 = 35.5$ yr, $p_3 = 0.02$ for $T_3 = 45.5$ yr and $p_4 = 0.01$ for $T_4 = 55.5$ yr [182], so the average generation time can be estimated as

$$\langle T \rangle = \sum_{i=1}^{N} p_i T_i, \tag{2.12}$$

and this leads to the generation time that will be used in the simulations of Chapter 5, namely $\langle T \rangle \approx 32$ yr [182]. The average generation time for hunter-gatherers is presumably similar to this value, so we will use it for both farmers and hunter-gatherers (using a different value for each of both populations would require more complicated simulations, and we do not expect that they would lead to substantial changes in the results).

2.2.6. **Net fecundities of farmers ($R_{0,F}$) and hunter-gatherers ($R_{0,HG}$)**

When dealing with differential equations, such as the Fisher Eq. (1.9), net reproduction (i.e., the effect of births minus deaths) is usually logistic,

$$\left.\frac{\partial p_F(x,y,t)}{\partial t}\right|_g = F(p_F) = a_F p_F(x,y,t)\left(1 - \frac{p_F(x,y,t)}{p_{F\,max}}\right), \tag{2.13}$$

where subscript F denotes farmers, g denotes population growth (i.e., net reproduction), a_F is the initial growth rate of farmers, and $p_{F\,max}$ is their maximum population density (also called saturation density or carrying capacity). Ethnographic values of the initial growth rate a_F for 3 pre-industrial populations that settled in empty space have been reported in Refs. [51, 183], as follows.

(i) The island of Pitcairn, which is located 4,000 miles West of Chile and 1,400 miles Southeast of Haiti, during years 1790-1856 [184]. Those data yield $a_{F\,Pitcairn} = 0.02995 \pm 0.00119$ yr^{-1} [183].

(ii) The Bass Strait islands (between Australia and Tasmania) during years 1820-1945 [184]. The corresponding data yield $a_{F\,Bass} = 0.02626 \pm 0.00052$ yr^{-1} [183].

(iii) The islands of Tristan da Cunha (in the middle of the South Atlantic Ocean) during years 1892-1946 [185]. Those data yield $a_{F\,Tristan} = 0.02527 \pm 0.00032$ yr^{-1} [183].

Moreover, for the United States population during the 19th century, data reported by Lotka [186] yield $a_{F\,US} = 0.03135 \pm 0.00063$ yr^{-1} [183] (the effect of immigration from Europe was not subtracted, so it can be argued that the intrinsic rate could be somewhat lower).

Isern et al. [183] calculated that these 4 ranges yield an 80%-confidence level average of $a_F = 0.028 \pm 0.005$ yr^{-1}.

Now note that for low values of the population density p, i.e. at the conditions of the leading edge of the expansion front, Eq. (2.13) becomes $F(p_F) \approx a_F p_F(x, y, t)$. Then

$$\frac{\partial p_F(x, y, t)}{\partial t} \approx a_F p_F(x, y, t), \tag{2.14}$$

and therefore

$$p_F(x, y, t + T) \approx p_F(x, y, t) e^{a_F T}. \tag{2.15}$$

As mentioned in Sec. 1.2.5, in Chapter 5 we use a very simple model in which, following previous work [47, 139], net reproduction is given by

$$p_F(x, y, t + T) = R_{0,F} p_F(x, y, t), \tag{2.16}$$

where $R_{0,F}$ is the net fecundity of farmers, until the carrying capacity is reached, see the second line in Eq. (1.32). It is true that we could apply the solution to the logistic Eq. (2.13) instead of (2.16), but since we require that [47, 139]

$$R_{0,F} = e^{a_F T}, \tag{2.17}$$

the front speed is the same for both models [139], and we therefore expect similar results. Thus, in the simulations in Chapter 5 we use the simpler Eq. (2.16).

Using the average value of a_F above ($a_F = 0.028$ yr^{-1}), and the mean value in Sec. 2.2.5 for the generation time ($T = 32$ yr), we obtain that $R_{0,F} = 2.45$ for farmers. In the simulations in Chapter 5, we assume that hunter-gatherers do not experience net growth (i.e., $R_{0,HG} = 1$) because initially they are at their maximum density in all inhabitable cells and, even if some of them become farmers, they will still need space [9].

2.3. Fronts from reaction-diffusion models

Chapters 3-4 in this thesis use reaction-diffusion differential equations. A very simple example is the Fisher Eq. (1.9). We use two methods to determine wave-of-advance speeds from reaction-diffusion equations, namely analytical calculations and numerical integrations. We summarize both approaches in Secs. 2.3.1 and 2.3.2 below, respectively.

Finally, in Sec. 2.3.3 we describe the computational model used in Chapter 5. In contrast to Chapters 3-4, the model in Chapter 5 is *not* based on differential equations. Indeed, it is a discrete-time model, i.e., the relevant quantities are computed every time step equal to one generation. The reason for this difference is the following. If the diffusion approximation were valid, for a population initially in a small region, at later times a Gaussian distribution for the population density (or concentration) versus distance would be observed. This is a very well-known result from basic diffusion theory (see, e.g., Fig. 2.1 in Ref. [101]). Whereas for viruses (Chapters 3-4) this result has been never questioned, and virus diffusion coefficients have been measured in many experimental studies (see Sec. 2.1.1), for pre-

industrial human populations (Chapter 5) it has been observed that the mobility data (sometimes called dispersal kernels) do not fit to a Gaussian [187, 183]. Thus, the diffusion approximation beaks down for pre-industrial human dispersals. But diffusion is only an approximation to more precise equations (see Sec. 1.2.2), which are called integro-difference equations (they have the form of Eqs. (1.12)-(1.13)). Indeed, the front speeds for humans obtained from reaction-diffusion equations make substantial errors, relative to the corresponding integro-difference equations [183]. Thus, in Chapter 5 we do not use differential equations but simple simulations that mathematically correspond to integro-difference equations [139], i.e. equations similar to Eqs. (1.12)-(1.13). More details are given in Sec. 2.3.3 below.

2.3.1. **Analytical calculations**

In contrast to the Fisher Eq. (1.9), which describes a single population, we will deal with sets of differential equations that describe 3 populations, namely viruses, uninfected and infected bacteria (in Chapter 3) or viruses, uninfected and infected tumor cells (in Chapter 4). We have already seen some examples of such equations in Sec. 1.2, e.g. Eqs. (1.19)-(1.21) or Eqs. (1.22)-(1.24). A front or travelling wave is defined as a population profile that travels with constant shape and speed. In order to accomplish an analytical solution for the front speed, three assumptions are made. (i) The population density p of the expanding species at the leading edge of the front is very small, therefore the equation describing its dynamics can be linearized (an example of this has been already given above, namely Eq. (2.14) is the linearized version of Eq. (2.13)). (ii) For $t \to \infty$ and $r \to \infty$ (where $r = \sqrt{x^2 + y^2}$ is the radial coordinate), the front of the expansion can be assumed planar at the local scale, which implies that the only non-vanishing spatial derivative is that along the radial direction, thus $\frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} \approx \frac{\partial^2 p}{\partial r^2}$. (iii) In the leading edge of the front, i.e. for $z \to \infty$ (where $z = r - ct$ is the co-moving coordinate and c is the front speed), the ansatz $p(z) = \tilde{p}e^{-\lambda z}$ for the population that expands its range (e.g., viruses) is assumed. Similarly, for a population that contracts its range (e.g., uninfected cells) or that has the shape of a pulse (e.g., infected cells) we assume that $q(z) = \tilde{q}\left(1 - e^{-\lambda z}\right)$ [188]. For an example of the shapes of travelling waves in these three cases, see Fig. 1.1, right.

The Fisher Eq. (1.9), the HRD Eq. (1.16), sets of reaction-diffusion equations such as (1.19)-(1.21) due to Yin and McCaskill, etc., can be solved by applying the three assumptions (i)-(iii) in the previous paragraph, although not always explicitly. We can derive the following equations from assumption (iii) above,

$$\begin{aligned} \frac{\partial p}{\partial t} &= \lambda c p(r,t), & \frac{\partial^2 p}{\partial t^2} &= (\lambda c)^2 p(r,t), \\ \frac{\partial p}{\partial r} &= -\lambda p(r,t), & \frac{\partial^2 p}{\partial r^2} &= \lambda^2 p(r,t). \end{aligned} \tag{2.18}$$

Then, according to marginal stability analysis [189], the front speed is given by

$$c = min_{\lambda>0}[c(\lambda)], \tag{2.19}$$

where $c(\lambda)$ is the so-called characteristic equation.

We illustrate this procedure for a simple example, namely the Fisher Eq. (1.9). First, applying assumption (i) above, we linearize Eq. (1.9). As explained in Sec. 2.2.6, the logistic Eq. (2.13) is linearized into Eq. (2.14). Thus, the Fisher Eq. (1.9) is linearized into Eq. (1.11). Next, we apply assumption (ii) so that we can replace the first term in the right-hand side of Eq. (1.11) by $D\frac{\partial^2 p}{\partial r^2}$. Finally, we apply assumption (iii), i.e. Eqs. (2.18) and obtain $D\lambda^2 - c\lambda + a = 0$. This is the characteristic equation for this simple case (i.e., for the Fisher Eq. (1.9)). Finally, we solve this second-order equation and require that λ is real, which immediately leads us to $c > 2\sqrt{aD}$. Thus, the minimum speed is $2\sqrt{aD}$ and, finally, Eq. (2.19) implies that the speed of the front is $c = 2\sqrt{aD}$. This is the derivation of the Fisher speed, Eq. (1.10). The Fisher speed is applied in Chapter 4, Eq. (4.24).

Chapters 3 and 4 apply this method to more complicated cases, namely to derive the speed of virus infections from their respective sets of differential equations. In these cases, however, it is not possible to obtain an explicit result as for the Fisher equation above, but we can still find the speed (given by the minimum of a function) numerically.

2.3.2. **Numerical integration**

Numerical integration makes it possible to obtain the speed of reaction-diffusion fronts, but only for specific numerical values of all of the parameters appearing in the reaction-diffusion equations. This approach is very useful to check the analytical calculations of the front speed described in the previous subsection. Moreover, numerical integrations make it possible to visualize the shape of travelling waves (for which seldom analytical equations can be derived). The so-called finite-difference method is usually applied to solve partial differential equations. It is based on approximating them to discrete finite-difference equations. In this approach, the following approximation for the first derivative of an arbitrary function f is used,

$$f'(x_0) \approx \frac{f(x_0 + h) - f(x_0)}{h}. \tag{2.20}$$

Using superscripts for time and subscripts for space, we have for the population density p

$$\frac{\partial p}{\partial t} \approx \frac{p_j^{n+1} - p_j^n}{\Delta t}, \qquad \frac{\partial^2 p}{\partial t^2} \approx \frac{p_j^{n+1} - 2p_j^n + p_j^{n-1}}{(\Delta t)^2},$$
$$\frac{\partial p}{\partial r} \approx \frac{p_{j+1}^{n+1} - p_j^{n+1}}{\Delta r}, \qquad \frac{\partial^2 p}{\partial r^2} \approx \frac{p_{j+1}^{n+1} - 2p_j^{n+1} + p_{j-1}^{n+1}}{(\Delta r)^2}. \tag{2.21}$$

Note that we evaluate the spatial derivatives at time $n + 1$, rather than n. This is called the implicit procedure, and it leads to correct results withouth need to use so small space and time intervals as the explicit procedure (in which the spatial derivative would be evaluated at time n) [190]. When the implicit procedure is applied to a reaction-diffusion equation, generally we obtain, for each time step $n + 1$, a set of linear equations that have the so-called tridiagonal form [190],

$$\begin{bmatrix} b_1 & c_1 & & & 0 \\ a_2 & b_2 & c_2 & & \\ & a_3 & b_3 & \ddots & \\ & & \ddots & \ddots & c_{J-1} \\ 0 & & & a_J & b_J \end{bmatrix} \begin{bmatrix} p_1^{n+1} \\ p_2^{n+1} \\ p_3^{n+1} \\ \vdots \\ p_J^{n+1} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_J \end{bmatrix}, \tag{2.22}$$

where p_i^{n+1} is the population density at distance $r_i = i\,\Delta r$ from the origin (i =1,2,3,..., J), and the spatial upper bound J is related to the size of the system L as $J = \frac{L}{\Delta r}$. The coefficients $d_1, d_2, \ldots, d_J$ depend on the population density at the two previous times, namely p_1^n and p_1^{n-1}. When dealing with an HRD equation (such as Eq. (1.16)) for a single population we apply, as initial conditions, that $p_j^1 = p_j^0$ is different from zero only in a central region (namely, that from which the population expands its range). If there are several species (e.g. viruses, uninfected cells and infected cells), we have a set of differential equations, one for each species. For each of them, we have a matrix with the form of Eq. (2.22) and the coefficients a_i, b_i, c_i, d_i also depend on the concentrations of the other species (at times $n-1$ and n). When dealing with several species, their initial conditions are those appropriate for the experimental setup. For example, for virus plaques we can assume the following initial conditions: (i) there are viruses only in a central region (where they have been inoculated); there are uninfected cells everywhere; and (iii) there are not infected cells anywhere [21].

Sets of linear equations of the form (2.22) can be easily solved with the *Tridag* routine in Fortran [190], and we have done so in our software codes. In this way, for each value of time $(n+1)$, we obtain the population density p_i^{n+1} at consecutive distances from the origin (i =1,2,3,..., J). Thus, numerical integration allows us to visualize the shape of travelling waves. An example can be seen in Fig. 1.1, right. Also, with this method we can calculate the front propagation speed easily, because by linear interpolation we can easily find the position where each travelling wave attains any given value, at each time step. For example, if a population attains, e.g., half of its maximum possible density at points i and j (with distances $r_i = i\,\Delta r$ and $r_j = j\,\Delta r$ to the origin) at times t_i and t_j, respectively, a rough estimate of the front speed is simply $\frac{r_j - r_i}{t_j - t_i}$. In our software codes, we obtain more precise estimates by fitting a linear regression to the values of the distances (r) versus times (t) at which the population density is, e.g., half the maximum possible value (i.e., half the carrying capacity). Since the shape of travelling waves is constant (at large enough times), the speed would be the same if we considered, e.g., a quarter or a tenth of the carrying capacity.

2.3.3. **Computational simulations**

As explained at the beginning of Sec. 2.3 (paragraph 2), reaction-diffusion equations do not provide an accurate description for human populations (Chapter 5). The reason is that in contrast to, e.g., viruses or tumor cells (Chapters 3-4), diffusion theory breaks down for humans (Chapter 5). This is because, contrary to the predictions of diffusion theory [101], for humans the number density of a population (initially in a small region) as a function of distance is given by a function (called the dispersal kernel) which does not have the shape of a Gaussian [187]. Therefore, using the diffusion approximation leads to substantial errors for the speed of fronts, of up to about 50% [183]. This is why for human populations (Chapter 5) we prefer to use integro-difference equations (i.e., equations with the form of Eqs. (1.12)-(1.13)). We stress that the reaction-diffusion approximation is obtained from integro-difference equations assuming small enough dispersal distances and times (Sec. 1.2.2), so

integro-difference equations provide a more precise description than differential equations. Another difference is that in Chapter 5 (human populations) we will perform computational simulations on a real map of Europe (i.e., including seas and mountains), whereas in Chapters 3 and 4 (virus systems) we consider homogeneous space.

In Chapter 5 we use computational simulations, rather than integro-differential equations, although both approaches are equivalent, because we are interested in the shape of a genetic cline (i.e., in the frequency of a genetic marker as a function of position), and this problem cannot be solved analytically but only numerically.

The details of the simulation applied in Chapter 5 are given within the same Chapter (Sec. 5.8.5), and therefore this section only summarizes its main features.

We run our simulations on a Cartesian grid of square cells of 50×50 km^2 each. They are classified as inland, coastal, mountain, or sea cells. The entire grid covers the whole European continent and the Near East. The cell side is 50 km, because this is the characteristic distance moved by preindustrial farmers per generation, as estimated from observed mobility and persistence data (in Sec. 2.2.2). This model follows a dispersion-interaction-reproduction scheme, which allows us to study the evolution of a genetic haplogroup associated with early farmers. In agreement with archaeological evidence, we initially set the Neolithic population (farmers) in the Near East (as a representative origin of the spread, we use the cell in Syria with the oldest Neolithic site of the culture that spread into Europe). Initially, the rest of the inhabitable cells are populated by Mesolithic hunter-gatherers at their saturation density (see Sec. 2.2.4 for population density values). In agreement with ancient DNA data (Chapter 5), some of the initial farmers have the mitochondrial haplogroup K, but none of the HGs has haplogroup K. At each iteration (or generation), the computational model of Chapter 5 performs three sequential steps: population dispersal, population interaction and population growth. The three mechanisms are summarized below (see Sec. 5.8.5 for a more extended description).

Population dispersal: at each iteration, and for each inhabited cell, a fraction $p_e = 0.38$ (see Sec. 2.2.2) of the Neolithic population remains in the same cell. The rest of the farmers, $(1 - p_e)$, will migrate to other cells. If the origin cells is an inland cell, the remaining individuals are equally distributed among the four nearest neighboring cells, all of which are located at a distance of 50 km (mountain cells are not inhabitable, therefore if one neighbor is a mountain cell, the population that moves is distributed equally among the other three cells). If the origin cell corresponds to a coast cell, at least one of its neighboring cells will be a sea cell, thus the corresponding fraction of Neolithic individuals can travel across the sea and relocate with equal probability in other coast cells within a certain range (150 km, see Sec. 5.8.6). For a practical example of how dispersion works, see Fig. 2.5. In this example, initially we have $P_{F,0} = 100$ farmers at the source coast cell (marked by a red dot) surrounded by one cell of each type (mountain cell at the top, inland cell at right, coast cell at left, and sea cell at the bottom), as shown in Fig. 2.5, left. Of these $P_{F,0} = 100$ farmers, a constant portion p_e will remain at the site, i.e. $p_e P_{F,0} = 38$ farmers, see Fig. 2.5 right.

Then, the rules above imply that the rest of the farmers $(1 - p_e)$ will migrate to three neighboring cells, since the fourth cell (to the top) is a mountain cell, and thus not inhabitable. The three possible travel directions are to the right and to the left (by land travel) and to the bottom (by sea travel). Thus, we obtain that $(1 - p_e)P_{F,0}/3 = 20.\hat{6} \approx 21$ farmers will move to the left cell and 21 farmers will

move to the right cell (see Fig. 2.5, right). Because the sea cell cannot be inhabited, the population that would travel there is equally distributed into all possible destinations according to the rules of sea travel. In this case we assumed a maximum sea travel range of 150 km, and thus there are 5 coast cells that can be reached by sea. Each of the 5 coast destinations reached by sea travel will receive $20.\hat{6}/5 \approx 4$ farmers in this example. Note that the cell to the left of the origin cell will receive 25 individuals, 21 through land travel and 4 through sea travel. Note as well that the number of farmers travelling to the bottom (20 farmers, which travel by sea) is different to that travelling to the left or to the right (21 farmers). However, in Chapter 5 we will use population densities rather than numbers, i.e. we will compute using non-integer numbers ($20.\hat{6}$ in this case). In future work, it would be of interest to perform simulations using integer numbers, so that the effects of stochasticity could be analyzed.

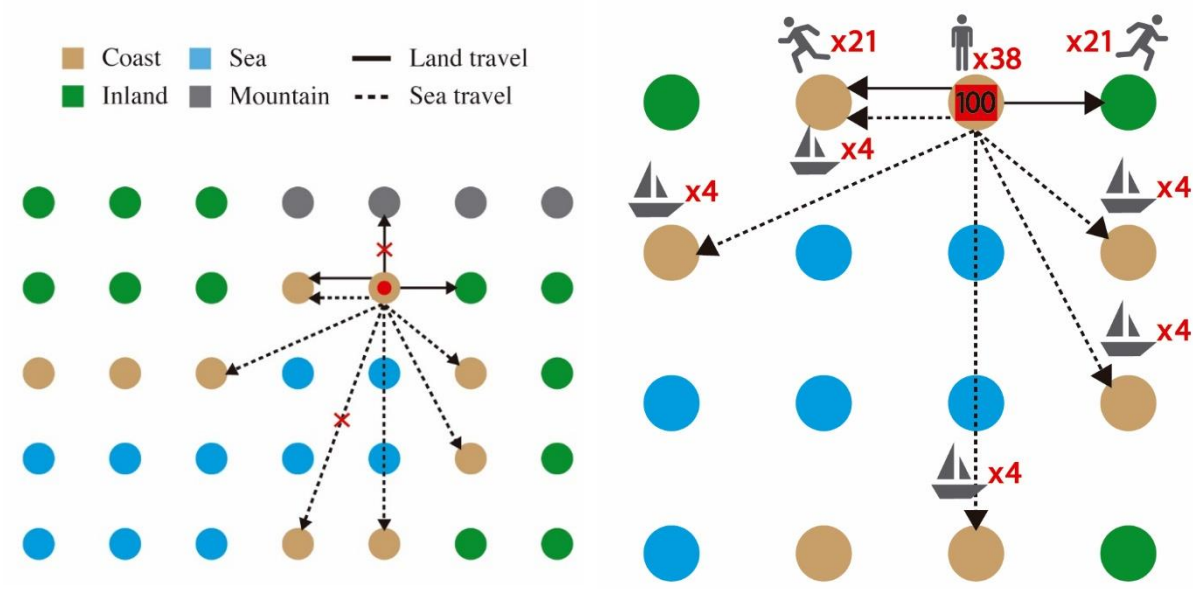


Figure 2.5 Left: Example of the model used in Chapter 5. The population in the initial coast cell (red) can travel (i) by land to the right (inland cell) and to the left (coast cell), but no to the top (mountain cells are not inhabitable), (ii) by sea because the bottom cell corresponds to a sea cell (four possible destinations are at distances equal or lower than 150 km, which is the maximum sea travel distance, see Sec. 5.8.6). Right: Distribution of the 100 individuals at the initial coast cell (red) one iteration later: 38 individuals remain at the same spot and 62 individuals migrate (42 by land and 20 by sea travel).

Population interaction: the computational model in Chapter 5 considers that the Neolithic transition could have been a mix of demic and cultural diffusion processes. Vertical cultural transmission, i.e. cross-mating between farmers and hunter-gatherers, is computed at each iteration and inhabited cell using Eqs. (1.34)-(1.35). Those equations show that a fraction (ruled by the interbreeding parameter η) of the Mesolithic hunter-gatherer (HGs) population will mate with farmers (Fs) and become part of the Neolithic community (ethnographic data indicate that HGs often become Fs, but in contrast Fs very rarely become HGs [52]). This incorporation of HGs into populations of Fs is an important feature of the model, because their children can inherit HG genetic markers and this leads to a decrease of the Neolithic haplogroup K in populations of Fs with increasing distance from the origin of the Neolithic spread (i.e., to a genetic cline).

Population growth: unless the farmer population density (in a given cell) has reached its saturation value, it will increase by a factor $R_{0,F}$ (see Sec. 2.2.6). For simplicity, we assume that hunter-gatherers will not increase their population density, even if it has diminished due to cultural transmission. This seems a reasonable approximation, because cultural transmission yields new farmers which will use part of the cell space. In a more complicated model, we could include hunter-gatherer net reproduction, but we consider it likely that this would yield similar results, because the increase in farmer density due to cultural transmission is presumably small, compared to farmer net reproduction. For example, in Chapter 5 we will find that only about 2% of farmers interbred with hunter-gatherers (or acculturated them). This means that cultural transmission leads, for 100 initial farmers, to 102 farmers a generation later. But ethnographic data indicate a net reproduction rate of about $R_{0,F} = 2.45$ (Sec. 2.2.6). Thus, net reproduction leads, for 100 initial farmers, to 245 farmers a generation later. Therefore, the effect of population growth (on population number) is surely larger than that of cultural transmission.

PART II

Results

3. Front propagation speeds of T7 virus mutants[2]

Abstract We propose a new reaction-diffusion model with an eclipse time to study the spread of viruses on bacterial populations. This new model is both biologically and physically sound, unlike previous ones. We determine important parameter values from experimental data, such as the one-step growth. We verify the proposed model by comparing theoretical and experimental data of the front propagation speed for several T7 virus strains.

Keywords biophysics, front propagation, mathematical model

3.1. Introduction

Bacterial viruses or bacteriophages (literally `eaters of bacteria') infect and replicate within bacteria. Right after their discovery, phages were used as an early form of biotechnology to fight bacterial pathogens. Nowadays, drug-resistant strains for many bacteria have appeared and this has led to a revived interest in this kind of therapy [191]. Moreover, these viruses are among the most common and diverse entities in the biosphere, so it is important to attain a better and more accurate knowledge of their dynamics. Understanding the speed of virus infection fronts is also important in the context of cancer treatment [45].

It is possible to see with the naked eye how the spreading dynamics of viruses works in a medium of susceptible host bacteria. When a small quantity of phages is inoculated into a tiny, central region of liquid agar with host cells (bacteria in our case), the continuous replication and diffusion of viruses lead to an enlarging dark region, composed of dead cells. Such a region of lysed (i.e. dead) cells, surrounded by unlysed cells, is called a plaque. The growth process starts when a free virus diffuses into a host bacterium, adsorbs on its surface, injects its DNA into it, replicates within and finally (after a certain time) the bacterium dies and expels a new generation of viruses. The progeny viruses diffuse to surrounding host cells, and the cycle repeats again. The propagating front has a well-characterized speed, typically less than a millimeter per hour, which has been measured experimentally, and for which we try below to get a realistic and accurate reaction-diffusion model.

Numerous models of phage plaque enlargement have been proposed. The oldest and simplest one is due to Koch [13], who suggested that the diffusion speed was proportional to $\sqrt{\frac{D}{\tau}}$, where D is the diffusion coefficient and τ the phage latent period (i.e., the time during which bacteriophages are inside cells, and thus not moving). By incorporating additional kinetic parameters, Yin and McCaskill

[2] This Chapter is an exact transcription of the contents of the following paper (please find a copy of the published version in Appendix B): de Rioja VL, Fort J, Isern N. Front propagation speeds of T7 virus mutants. *Journal of Theoretical Biology* **385**, 112–118 (2015). DOI: 10.1016/j.jtbi.2015.08.005.

[14] constructed a reaction-diffusion system and obtained the speed of travelling-wave solutions. Later You and Yin [144] supported the previous idea of an existing travelling-wave solution through numerical simulations of the same problem. However, the models due to Yin and co-workers [14, 144] lead, for parameter values derived from independent experiments, to speeds much faster than the experimental ones [14, 19]. It was then realized that the delay time or latent period (i.e., the time interval during which a virus is inside a cell and thus does not move) delays virus diffusion, and that this important effect could explain the slowness of the experimental speeds [4]. By solving the problem numerically, good agreement with experiment was attained (without fitting any parameter values) [4]. However, the equations were not fully understood from a biological viewpoint, as we shall explain below. Later Ortega-Cejas et al. [115] obtained some approximate but explicit formulas for the front speed based on the model in Ref. [4]. Amongst more recent models, Amor and Fort [21] proposed a new improved set of equations which satisfactorily explained the observed speeds of VSV (Vesicular Stomatitis Virus) infections, but still with some terms lacking a clear biological interpretation.

Using various bacteriophage T7 mutants in a growing plaque on *E. coli* host bacteria, Yin measured experimentally [15] the radial propagation speed for plaques of three mutant T7 virus strains (namely, p001, p005 and the wild type), finding different speeds depending on the type of mutant. These are the experiments that we want to explain.

On the basis of the model for VSV infections [21], we rewrite the equations carefully, so that they acquire full biological and mathematical meaning, and we apply them to T7 strains. With this new model we obtain a good agreement with the experimental results in Ref. [15], without requiring the use of any free or adjustable parameters.

In this paper, we introduce a new reaction-diffusion set of equations to explain the existing experimental data on the growth of T7 plaques on bacteria. In Sec. 3.2 we present the new time-delayed model and we discuss why our modifications are reasonable. Section 3.3 is devoted to estimations of the necessary parameter values from independent experiments. In Sec. 3.4, the results are compared with experimental data for the propagation speed of three strains of the T7 virus, and Sec. 3.5 presents a simplification of the model yielding similar results. In Sec. 3.6 we compare to other time-delayed models. Finally, Sec. 3.7 is devoted to final conclusions, with particular attention to the model features and how the results are improved over previous models.

3.2. Reaction-diffusion model

We model the spatial dynamics of T7 mutants infecting host cells by considering interactions between three species: viruses (V), uninfected bacteria (B) and infected bacteria (I). Those processes can be described schematically as

$$V + B \overset{k_1}{\rightarrow} I \overset{k_2}{\rightarrow} Y \cdot V, \tag{3.1}$$

where k_1 is the adsorption rate, k_2 the death rate of infected bacteria, and *Y* (yield or burst size) is the number of new viruses released per lysed host bacteria. These three parameters (k_1, k_2 and *Y*) depend on the mutant strain considered.

The experiment on which this theoretical work is focused [15] was conducted in agar (so that host bacteria are immobilized) and cells were initially in the stationary phase, i.e. with bacterial growth and death in balance (so that the number density of live bacteria does not change appreciably before viruses arrive). Viruses can move and adsorb on host bacteria, infecting cells and producing new viruses.

Some previous models have the drawback of assuming logistic dynamics, namely [4, 21, 115]

$$\frac{\partial[I](r,t)}{\partial t} = -k_2[I](r,t)\left\{1 - \frac{[I](r,t)}{[I]_{max}}\right\}, \tag{3.2}$$

in the absence of uninfected cells (i.e., if all cells are initially infected). $[I]$ in Eq. (3.2) is the concentration of infected hosts, $[I]_{max}$ their maximum concentration, r the distance from the inoculation point, and t the time.

Let us define "free space" as the fraction of space not occupied by infected cells, relative to the maximum possible value that can be occupied by them, i.e. $1 - \frac{[I](r,t)}{[I]_{max}}$. Equation (3.2) describes well the one-step growth experiment (see Fig. 1 in [19]) but has no biological meaning. Indeed, it assumes that the death rate of infected cell is proportional not only to the concentration of infected cells $[I]$ (which is reasonable), but also to the free space (term within brackets). Thus, we propose to replace this equation by taking into account the eclipse time τ between adsorption and the onset of the release of the virus progeny. Therefore, in the absence of adsorption we propose to replace Eq. (3.2) by

$$\frac{\partial[I](r,t)}{\partial t} = -k_2[I](r,t-\tau). \tag{3.3}$$

Note that we do not assume that all cells die at the same time after infection. That assumption is made in the perfect delay model [147], which makes use of $[V](r,t-\tau)[B](r,t-\tau)$ instead of $k_2[I](r,t-\tau)$ (as we will see in Sec. 3.6 in detail). But the perfect delay model disagrees with biological experiments (because one-step experiments do not display a vertical step, see Fig. 3.1). In contrast, Eq. (3.3) does not represent a perfect vertical step, but a gradual increase after an eclipse time τ. The model we present below is an alternative to the exponential, non-delayed model [i.e., a term $k_2[I](r,t)$] and the perfect delay model [i.e., a term $[V](r,t-\tau)[B](r,t-\tau)$] and is more realistic than both extreme models.

In the presence of adsorption, the model we propose is thus described by (see Sec. 3.9):

$$\frac{\partial[I](r,t)}{\partial t} = k_1[V](r,t)[B](r,t) - k_2[I](r,t-\tau), \tag{3.4}$$

$$\begin{aligned}\frac{\partial[V](r,t)}{\partial t} + \frac{\tau}{2}\frac{\partial^2[V](r,t)}{\partial t^2} &= D_{eff}\frac{\partial^2[V](r,t)}{\partial r^2} + F(r,t) - \frac{\tau}{2}k_1[V](r,t)\frac{\partial[B](r,t)}{\partial t} \\ &\quad - \frac{\tau}{2}k_1[B](r,t)F(r,t) + \frac{\tau}{2}k_2 Y\frac{\partial}{\partial t}\{[I](r,t-\tau)\},\end{aligned} \tag{3.5}$$

$$\frac{\partial [B](r,t)}{\partial t} = -k_1[V](r,t)[B](r,t), \tag{3.6}$$

where $[V]$ and $[B]$ are the concentration of viruses and uninfected bacteria respectively, and D_{eff} is the effective diffusion coefficient of viruses (see next section). Bacteria do not diffuse because they are immobilized by the agar in this experiment. The virus growth function, $F(r,t)$, in Eq. (3.5) is

$$F(r,t) = -k_1[V](r,t)[B](r,t) + k_2 Y[I](r,t-\tau). \tag{3.7}$$

In this model [Eqs. (3.4)-(3.7)], the time derivative $\frac{\partial}{\partial t}$ represents the change of the population number over time and the second space derivative $\frac{\partial^2}{\partial r^2}$ is related to the diffusion through space. Terms proportional to k_1 account for the decay of viruses [Eqs. (3.5) and (3.7)] and host bacteria [Eq. (3.6)] and the creation of infected cells [Eq. (3.4)], as a result of the infection process (note that these terms are the same as Eqs. (9) and (11)-(13) in Ref. [21]). Infected cells also decay following their own rate of death k_2 [Eq. (3.4)], and as shown by Eq. (3.1), for each dead cell the viruses increase their number *Y* times [Eqs. (3.5) and (3.7)]. The terms proportional to τ in Eq. (3.5) are second-order corrections (see Sec. 3.9), they were applied already in Ref. [21], and they take care of the time delay due to the fact that viruses spend a time τ inside cells before the new generation disperses away [4]. As mentioned above, the main drawback of Ref. [21] is studying the death of the infected cells from a logistic equation, which has no biological sense.

Therefore, here we present a new model with two main effects: (i) the second-order correction that has been shown to be fundamental to describe time-delayed biological fronts [4, 21, 51] [i.e., the terms proportional to τ in Eq. (3.5)], and also (ii) a biologically meaningful description of the death process, Eq. (3.3) [instead of logistic growth dynamics, Eq. (3.2)].

Other authors have also described the death process through including an eclipse time with terms proportional to concentrations at $t-\tau$, rather than a logistic function [147, 146]. However, those models do not include any second-order terms, i.e. any diffusive delay (effect (i) in the previous paragraph), which is necessary to take proper account of the fact that viruses do not move during a time interval τ, because they are inside the infected cells. The death of infected cells is also described in Refs. [147, 146] differently than in our model (in Sec. 3.6 we discuss this in more detail and compare the models and experiments).

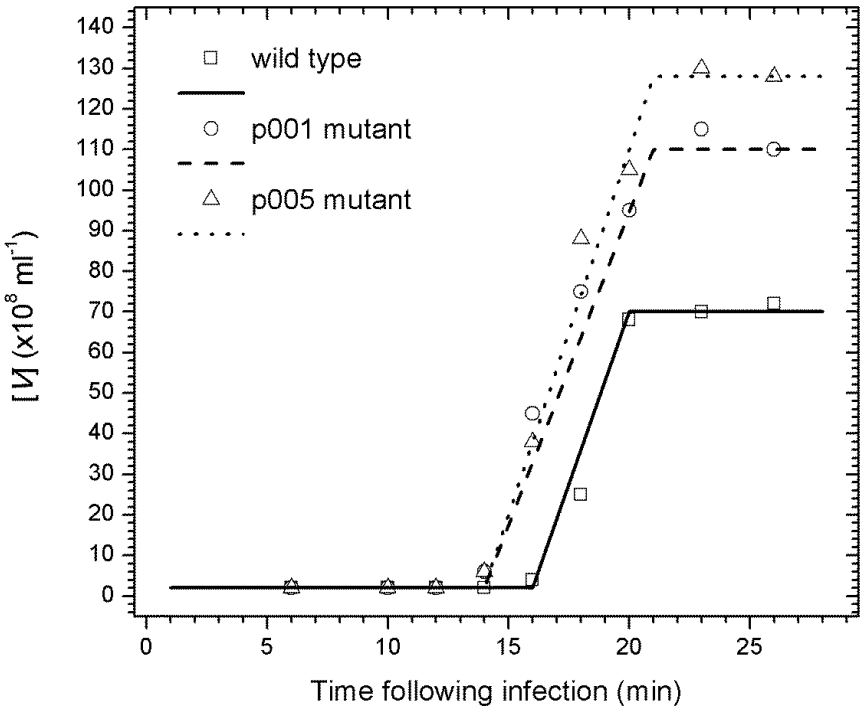


Figure 3.1 One-step growth curves of T7 mutants adapted to the model in this paper. Experimental data (□ for the wild T7, ○ for the p001 mutant and △ for the p005 mutant) have been obtained from Fig. 3 in Ref. [15]. Full, dashed and dotted lines correspond to the fits for the wild type and p001 and p005 mutants, respectively.

We introduce dimensionless variables to simplify the analysis. Let B_0 be the initial concentration of bacteria, then $\bar{B} \equiv [B]/B_0$, $\bar{V} \equiv [V]/B_0$, $\bar{I} \equiv [I]/B_0$, $\bar{t} \equiv k_2 t$ and $\bar{r} \equiv r\sqrt{k_2/D_{eff}}$ are the new dimensionless variables, and $\bar{\tau} \equiv k_2\tau$ and $\kappa \equiv k_1 B_0/k_2$ the new dimensionless parameters. The aim is to find the speed of travelling-wave solutions which satisfies the set of differential equations (3.4)-(3.6). These become single-variable differential equations by using the co-moving coordinate $\bar{z} = \bar{r} - \bar{c}\bar{t}$. $\bar{c}$ (positive) is the dimensionless wave front speed and is related to the dimensional speed *c* by $\bar{c} = c/\sqrt{k_2 D_{eff}}$. Following previous work [4, 14, 21], we assume that the concentrations at the leading edge of the propagation front ($z \to \infty$) are $(\bar{V}, \bar{B}, \bar{I}) = (\epsilon_V, 1-\epsilon_B, \epsilon_I) \approx (0,1,0)$, where $\epsilon = (\epsilon_V, \epsilon_B, \epsilon_I) = \epsilon_0 \cdot e^{-\lambda\bar{z}}$. For non-trivial solutions to exist, the determinant of the matrix corresponding to the linearized model must be zero. This leads us to the following characteristic equation

$$\left(1-\frac{\bar{\tau}\bar{c}^2}{2}\right)\bar{c}\lambda^3 + \left[e^{-\lambda\bar{c}\bar{\tau}} - \bar{c}^2\left(1+\frac{\bar{\tau}}{2}e^{-\lambda\bar{c}\bar{\tau}}\right)\right]\lambda^2 + \left[\kappa\left(\frac{\bar{\tau}}{2}\kappa - 1 + \frac{\bar{\tau}}{2}Ye^{-\lambda\bar{c}\bar{\tau}}\right) - e^{-\lambda\bar{c}\bar{\tau}}\right]\bar{c}\lambda + \kappa e^{-\lambda\bar{c}\bar{\tau}}\left[\frac{\bar{\tau}}{2}\kappa - 1 - Y\left(\frac{\bar{\tau}}{2}\kappa - 1\right)\right] = 0. \quad (3.8)$$

It is known that, according to marginal stability analysis [189], the propagation front moves with the minimum possible speed. Therefore,

$$\bar{c} = min_{\lambda>0}[\bar{c}(\lambda)], \quad (3.9)$$

where $\bar{c}(\lambda)$ is given implicitly by equation (3.8).

3.3. Parameter values

We have a new time-delayed model which depends on various parameters. It is necessary to estimate their values from experiments different from the front-speed experiments that we want to explain. The front propagation speed depends on the viral diffusivity D_{eff}, the average yield *Y*, the kinetic parameters k_1 and k_2, the host concentration B_0 and the eclipse time τ. Since we aim to explain the experimental data in Ref. [14], these parameters must be determined for strains of the T7 virus and *E. coli* bacteria.

Yin and co-workers noted that the diffusion coefficient *D* of viruses in agar must be corrected by the fact that host bacteria adsorb the viruses, and this leads to more tortuous paths for the viruses at high bacterial concentrations. As noted in previous work, the effective coefficient D_{eff} is therefore given by Fricke's law [4],

$$D_{eff} = \frac{1-f}{1+\frac{f}{x}}D, \tag{3.10}$$

where *f* is the initial concentration of bacteria relative to its maximum, i.e. $f = B_0/B_{max}$, and *x* stands as an approximation of the cells' shape. For spherical particles $x = 2$, while for *E. coli* it is more accurate to use $x = 1.67$ [4]. Note that the diffusivity coefficient *D* corresponds to viruses moving through agar in absence of bacteria ($f = 0$). T7 viruses are very similar to phage P22 in shape and size, thus we use the corresponding value $D = 4 \times 10^{-8}$ cm^2/s [4].

The rate of adsorption of viruses, k_1, was estimated from a separate experiment conducted in KCN, a substance that prevents viruses from reproducing. We have only one experimental value for the T7 virus, $k_1 = (1.29 \pm 0.59) \times 10^{-9}$ ml/min [4], corresponding to the wild strains. For the other mutants, we have not been able to find any reliable experimental value, thus we will use the same value of k_1 for all three strains. We will return to discuss this parameter in the next section.

Finally, the parameters τ, *Y* and k_2 are obtained from the so-called one-step growth experiments. They consist in measuring the concentration of viruses as a function of time for a given initial, homogeneous population of infected bacteria. Depending on the T7 mutant, the curves are different (see Fig. 3.1) and so will be the parameter values. Figure 3.1 allows us to obtain the necessary information from each mutant to estimate its value of τ, *Y* and k_2, as we next explain.

The eclipse phase of the one-step growth Fig. 3.1 corresponds to the stage between adsorption ($t = 0$) and the first release of viruses (i.e., the beginning of the rise in virus density). This interval of time (the eclipse time) is 16 minutes for the wild type and 14 minutes for the p001 and p005 mutants. Note that if we used higher values for τ, e.g. $\tau = 18$ min for the wild strain, for $\tau = 17$ min we would have $\frac{\partial [I](r,t)}{\partial t} = 0$ according to Eq. (3.3), whereas we must have $\frac{\partial [I](r,t)}{\partial t} \neq 0$ according to Fig. 3.1.

The average burst sizes *Y* differ significantly for the three mutants. They can be calculated as the quotient between the maximum and the initial concentration of viruses, i.e. $Y = \frac{V_{max}}{V_0}$, where according to Fig. 3.1 $V_0 = 2 \times 10^8$ ml-1 for the three kinds of mutants. Inserting the data in Fig. 3.1, we obtain the yields $Y = 34.5$ for the wild type, $Y = 56.5$ for the p001mutant, and $Y = 65$ for the p005 mutant. As we shall see, these higher productivities of new generations for the two mutants

result in faster infections (relative to the wild type). The three yields above have been obtained for cells in agar-immobilized microcolonies containing many cells. As noted by Yin and McCaskill [14], such yields are substantially lower than the typical yield for an isolated cell under optimal conditions ($Y \approx 200$). Yin and McCaskill suggested that this difference may be due to a number of factors, such as inherently lower yields per cell when immobilized in agar, premature lysis or inhibition due to the death of adjacent cells, high multiplicities of adsorption required for host infection, readsorption of newly released viruses on cell fragments, etc. [14]. If we used the yield for an isolated cell, we would have to incorporate additional terms to include other possible kinds of death and interactions in our mathematical model. However, the measured experimental values of the burst size quoted above (for cells in agar-immobilized microcolonies containing many cells) implicitly include these possible interactions.

The rate of death of infected bacteria k_2 may be understood as the reproduction of viruses, because viruses replicate as bacteria die. For $t < \tau$ there are no new viruses (see Fig. 3.1), so no infected cells have died yet and thus $[I](t) = I_0$. For $t \geq \tau$ Eq. (3.3) yields $d[I] = -k_2 I_0 dt$. Because each infected cell produces *Y* viruses, $d[V] = -Y d[I] = k_2 Y I_0 dt = k_2 V_{max} dt$. Therefore, the slope of each straight line in Fig. 3.1 is $k_2 V_{max}$, and $k_2 = \frac{V_{max} - V_0}{\Delta t \cdot V_{max}} \approx \frac{1}{\Delta t}$, where Δt is the time interval during which $[V]$ increases, also known as the rise period $\frac{1}{k_2}$. It is straightforward to estimate the values of k_2 from the figure, and they turn out to be 1/4 min^{-1} for the wild type and 1/6 min^{-1} for the p001 and p005 mutants.

It is important to remember that all of these parameters are known *a priori*, thus we do not use any free or adjustable parameters in our predictions.

Accordingly to Fig. 3.1, the average latent period is $\tau + \frac{1}{2k_2}$, where τ is the eclipse time.

For clarity, we mention that when all infected cells have died, no more viruses are produced to (Fig. 3.1, right side) and Eq. (3.3) obviously breaks down. Thus the general evolution equation we propose is

$$\frac{\partial [I](r,t)}{\partial t} = \begin{cases} -k_2 [I](r, t-\tau) \text{ if } [V] = V_{\text{max}} \\ 0 \quad \text{if } [V] = V_{\text{max}} \end{cases}, \tag{3.11}$$

where the second line is analogous to some approaches to single-species systems (see Eq. (9) in Ref. [139]). However this point is, in fact, unnecessary for the purposes of the present paper because the front speed is computed at the leading edge of the infection front, where $[V] \approx 0$ (Sec. 3.2). Obviously, in Eq. (3.11) the condition $[V] < V_{max}$ is equivalent to $[I] \neq 0$, and the condition $[V] = V_{max}$ is equivalent to $[I] = 0$.

3.4. Theory versus experiment

In this section we study the spatial dynamics of different T7 virus strains. The experimental data (black squares in Fig. 3.2) and their error bars were obtained in Ref. [15] for plaques where the concentration of nutrient was 10 g/l, which corresponds to $f = 0.2$ (see Ref. [14], pp. 1543-1544) and $B_{max} = 10^7$ ml^{-1} (see Ref. [14], Fig. 3a) thus $B_0 = 2 \times 10^6$ ml^{-1}.

The theoretical results will be calculated below with the parameters Y, k_2 and τ for each strain extracted from Fig. 3.1 (as detailed in Sec. 3.3), and the mean values of k_1 and D_{eff}. Because the value of k_1 is substantially more uncertain than those of other parameters [4], the corresponding error bars are obtained from the experimental range of k_1, namely $k_1 = (1.29 \pm 0.59) \times 10^{-9}$ ml/min [4].

The classical approach with no delay or eclipse time, due to Yin and McCaskill (triangles in Fig. 3.2), predicts speeds much faster than the experimental ones (black squares). This model by Yin and McCaskill [14] (with $k_{-1} = 0$, as noted in Ref. [192]) is the same as our model [Eqs. (3.4)-(3.7)] with $\tau = 0$, i.e.

$$\frac{\partial [I](r,t)}{\partial t} = k_1[V](r,t)[B](r,t) - k_2[I](r,t), \tag{3.12}$$

$$\frac{\partial [V](r,t)}{\partial t} = D_{eff}\frac{\partial^2 [V](r,t)}{\partial r^2} - k_1[V](r,t)[B](r,t) + k_2[I](r,t), \tag{3.13}$$

$$\frac{\partial [B](r,t)}{\partial t} = -k_1[V](r,t)[B](r,t). \tag{3.14}$$

The new model introduced in this paper (circles in Fig. 3.2) agrees better to the experimental data than the classical model Yin and McCaskill, for all three mutants. This improvement is clearly visible in Fig. 3.2, where we see that the results from the new model lie much closer to the experimental data than the classical model [14]. If we calculate the errors of the models versus the experimental data, the classical model by Yin and McCaskill has an average error of 75%, compared to only 10% for the new model presented here.

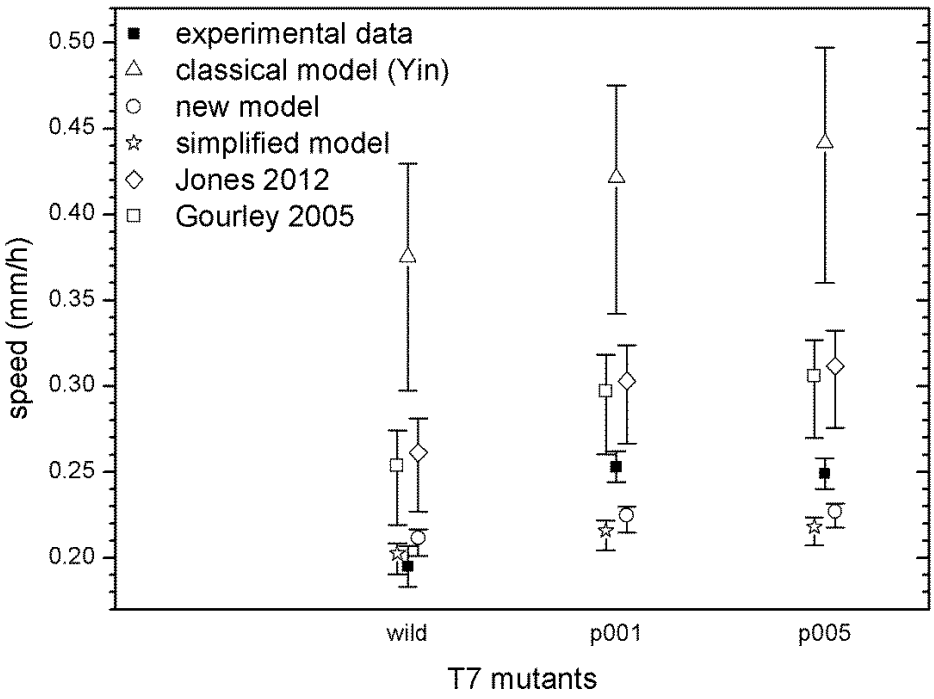


Figure 3.2 Front propagation speeds for T7 mutants (wild, p001 and p005). Black squares refer to experimental data and white symbols to the theoretical models: triangles for the classical Yin et al. model, circles for the new model, stars for the simplified model explained in Sec. 3.5, rhombuses for the model by Jones et al. (2012), and white squares for the model by Gourley et al. (2005) both from Sec. 3.6.

3.5. Simplified mathematical model

Our new model, Eqs. (3.4)-(3.7), yields a rather complex characteristic equation, Eq. (3.8), from which we compute the front speeds. In this section we derive a simplified expression leading to similar results. We proceed by removing each term and evaluating its contribution to the front speed, in order to ultimately keep only those terms that have a major contribution on the model results.

In this way, it is easy to see that all of the terms in Eqs. (3.4) and (3.6) are important to achieve a good result, but some terms in Eq. (3.5) are not. Hence, we just modify this equation.

On one hand, the expansion of $F(r,t)$ to second-order [the three last terms in Eq. (3.5)] introduces a small change on the results. We can neglect all reaction terms proportional to τ in this equation.

On the other hand, if we understand the right side of Eq. (3.5) as the diffusion term, plus the reaction term (plus second-order approximations), we can also neglect the adsorption of virus into bacteria, i.e. the term with k_1 in Eq. (3.7). Diffusion and creation of new viruses are thus the terms with major contributions to the front speed.

In this simplified model we can therefore replace Eq. (3.5) in our set by

$$\frac{\partial[V](r,t)}{\partial t}+\frac{\tau}{2}\frac{\partial^2[V](r,t)}{\partial t^2}=D_{eff}\frac{\partial^2[V](r,t)}{\partial r^2}+k_2[I](r,t-\tau), \tag{3.15}$$

Considering now the set composed by Eqs. (3.4), (3.6) and (3.15) we obtain a new characteristic equation,

$$\left[\lambda\bar{c}+\lambda^2\left(\frac{\bar{\tau}\bar{c}^2}{2}-1\right)\right]\left(\lambda\bar{c}+e^{-\lambda\bar{c}\bar{\tau}}\right)-\kappa Y e^{-\lambda\bar{c}\bar{\tau}}=0, \tag{3.16}$$

much simpler than the previous Eq. (3.8). The results of this model are shown as stars in Fig. 3.2. As it can be seen, the front speeds of the simplified model (stars) are always slightly slower than those found by the main model (circles). But the difference between the two models is only about 4% in all three cases. By comparing with the experimental data (black squares in Fig. 3.2, we see that the simplified model in this section [stars, Eq. (3.16)] is still much better than the classical one (triangles) in spite of being much simpler than the complete model in Sec. 3.2 [circles, Eq. (3.8)].

3.6. Comparison to other time-delayed models

Some other authors have also described the death process by considering concentrations at $t-\tau$, rather than a logistic function [147, 146]. However, as mentioned above, those models do not include the diffusive delay (i.e., second-order corrections), which is necessary because viruses do not diffuse when they are inside the infected cells. Another difference between our model and that in Ref. [147] is that the term $k_2[I](r,t-\tau)$ in our model is replaced by $k_1[B](r,t-\tau)[V](r,t-\tau)$. From a conceptual point of view, in our model the infected cells present at the system at time $t-\tau$ begin to die at time *t*, and do so gradually thereafter (with rate k_2). Thus not all cells die exactly at time t in our model, in agreement with the experimental data (Fig. 3.1). In contrast, according to the model in Ref. [147] all cells infected at time $t-\tau$ die exactly at time *t*, thus in the one-step experiment their model predicts a perfect step-like result, in disagreement with experimental data (Fig. 3.1). Thus we expect

the model by Jones et al. [147] to yield faster speeds than our model for two reasons: (i) they neglect the diffusive delay; and (ii) they neglect the fact that the death of some cells takes longer than τ after infection. Replacing D by D_{eff} (as explained in Sec. 3.3), the model by Jones et al. is (see Eqs. (2.2) in Ref. [147])

$$\frac{\partial[I](r,t)}{\partial t} = k_1[V](r,t)[B](r,t) - k_1[V](r,t-\tau)[B](r,t-\tau), \tag{3.17}$$

$$\frac{\partial[V](r,t)}{\partial t} = D_{eff}\frac{\partial^2[V](r,t)}{\partial r^2} - k_1[V](r,t)[B](r,t) + Yk_1[V](r,t-\tau)[B](r,t-\tau), \tag{3.18}$$

$$\frac{\partial[B](r,t)}{\partial t} = -k_1[V](r,t)[B](r,t). \tag{3.19}$$

Note that this is the same as the model by Yin et al. [Eqs. (3.12)-(3.14)], with $k_2[I](r,t)$ replaced by $k_1[B](r,t-\tau)[V](r,t-\tau)$. By following again the same method as in Sec. 3.2, we find that the characteristic equation for the model due to Jones et al. [147] is

$$\lambda^2 - \lambda\bar{c} + \kappa\left(Ye^{-\lambda\bar{c}\bar{\tau}} - 1\right) = 0. \tag{3.20}$$

Note that, in fact, the equation for $\frac{\partial[I](r,t)}{\partial t}$ above is not necessary to compute this speed, since $[I]$ does not appear in the other two equations of the model by Jones et al. [147].

In Fig. 3.2 (rhombuses) we have also included the predictions of the model by Jones et al., for the same parameter values used in our model. We see in Fig. 3.2 that their model [147] predicts faster speeds than our model, as expected. Moreover, they are faster than the experimental speeds. For the wild strain, our model is consistent with the experimental range. For the mutants p001 and p005, the mean speeds predicted by our model are also closer to the experimental means (although the error bars are larger for the model by Jones et al. [147], because the speed depends strongly on k_1).

There is one more time-delayed model of virus front spread, due to Gourley et al. [146]. It is very similar to that by Jones et al. [147], discussed above, but it assumes an additional, natural death process only for infected cells (with rate μ_I and unrelated to virus infection) that decreases the number density of infected cells after time τ by a factor $e^{-\mu_I\tau}$ [146]. Although no biological reason was given in Ref. [146] why an additional death process might affect only the infected cells (and not the uninfected ones), for completeness we next explore whether this model by Gourley et al. [146] changes the results of the model by Jones et al. [147] or not. Since this model by Gourley et al. [146] includes an additional death process for the infected cells, intuitively we expect that it could yield slower speeds than the model due to Jones et al. [147]. For the experimental conditions corresponding to the speeds that we analyze in the present paper (Fig. 3.2), the model proposed by Gourley et al. is (see Eqs. (1.1), (2.1) and (4.1) in Ref. [146])

$$\frac{\partial[I](r,t)}{\partial t} = k_1[V](r,t)[B](r,t) - e^{-\mu_I\tau}k_1[V](r,t-\tau)[B](r,t-\tau), \tag{3.21}$$

$$\frac{\partial[V](r,t)}{\partial t} = D_{eff}\frac{\partial^2[V](r,t)}{\partial r^2} - k_1[V](r,t)[B](r,t) + Ye^{-\mu_I\tau}k_1[V](r,t-\tau)[B](r,t-\tau), \tag{3.22}$$

$$\frac{\partial[B](r,t)}{\partial t} = -k_1[V](r,t)[B](r,t), \tag{3.23}$$

where we have neglected cell reproduction because in the experiments we want to explain, the cells were in the stationary growth phase before the arrival of viruses (as explained in Sec. 3.2). We have also neglected virus death because it is negligible [193]. We do not include diffusion of uninfected or infected cells because bacteria are immobilized in agar in these experiments (as mentioned in Sec. 3.2). By following again the same method, the characteristic equation in the model due to Gourley et al. is

$$\lambda^2 - \lambda\bar{c} + \kappa\left(Ye^{-\lambda\bar{c}\bar{\tau}}e^{-\mu_I\tau} - 1\right) = 0. \tag{3.24}$$

Again, in fact the equation for $\frac{\partial[I](r,t)}{\partial t}$ above is not necessary to compute this speed, since $[I]$ does not appear in the other two equations of the set. In Fig. 3.2 (plotted as white squares) we have also included the predictions of this model by Gourley et al. [146] using the experimental value $\mu = 0.4\text{h}^{-1}$ (from Fig. 7 in Ref. [194]). It is seen that its predictions are slower (as expected) but almost the same as those of the model by Jones et al. [147]. The speeds from both models are faster than the experimental ones.

Finally, it is worth to note that, in situations where infected cells exit that class due to some other form of interaction, it would be necessary to modify our model. For example, for an additional, natural death process with exponential dynamics for infected cells, the right-hand side in Eq. (3.4) would include an additional term $-\mu_I[I](r,t)$ and Eq. (3.5) should be modified accordingly.

3.7. Conclusions

We have proposed a new reaction-diffusion model with an eclipse time that satisfactorily explains the experimental results of T7 virus plaques on *E. coli*. This improvement over previous models have been attained by means of the careful modification of one of the evolution equations, which lacked biological significance.

Indeed, some previous models [4, 21, 115] assumed that the death rate of infected cells is proportional not only to their density, but also to the free space [Eq. (3.2)], which is not biologically reasonable. In contrast, the new model assumes that the death rate is proportional only to the density of infected cells, Eq. (3.3), which begin to die after a time lag τ, corresponding to the eclipse phase of Fig. 3.1. Thus our new model is more reasonable biologically. Moreover, our new model agrees reasonably well with experimental data, in contrast to the classical model without delay or eclipse time due to Yin and et al. [14, 144]. It is important to stress that Yin and co-workers already noted that their model was too fast for realistic parameter values, and only by fitting three parameters could it yield sufficiently slow speeds to agree with the experimental ones (Fig. 3 in Ref. [14]). In contrast, here we have not fitted any parameter but used realistic values, i.e. all parameter values we have applied have been obtained from independent experiments.

Other authors took into account the role of the eclipse or delay time τ, but only in the reactive and not in the diffusive process [147, 146], and they assumed the same eclipse time for all viruses. Those models yield faster speeds than the experimental ones. Also, we stress the importance of using

realistic terms to modelized the interactions, e.g. the death process of infected cells (i.e. the release of viral progeny).

Since the propagation of viruses is an active field of study in biophysics and medicine, having an underlying theory that is both mathematically and biologically sound is of special relevance. Furthermore, we have found that the results agree with experiments.

By means of the detailed analysis of a simple mathematical model, we have aimed to demonstrate that such physical models are able to explain the spatial dynamics of virus infections. Certainly, in order to have a more comprehensive understanding of the problem, extensive data gathering for several viruses and environments should be undertaken.

3.8. Acknowledgments

This work was partially funded by ICREA (Academia award) and the MINECO (projects SimulPast-CSD2010-00034, FIS-2009-13050 and FIS-2012-31307).

3.9. Time-delayed diffusion

In order to make this paper as self-contained as possible, here we include a brief derivation of Eq. (3.5). The derivation below (see Refs. [117, 51] for details) was originally proposed for human populations [51] and later applied to viruses [4, 21, 115].

During a time interval equal to the eclipse time τ (estimated from Fig. 3.1 in our case), the virus concentration changes both due to the reactive processes (3.1) and to dispersal. We first calculate the former change by using a Taylor series,

$$|[V](x,y,t+\tau)-[V](x,y,t)|_r = \tau\left.\frac{\partial[V]}{\partial t}\right|_r + \frac{\tau^2}{2}\left.\frac{\partial^2[V]}{\partial t^2}\right|_r + \cdots = \tau F + \frac{\tau^2}{2}\left.\frac{\partial^2[V]}{\partial t}\right|_r + \cdots \qquad (3.25)$$

where the subscript *r* denotes reactive processes, and $F([V]) = \left.\frac{\partial[V]}{\partial t}\right|_r$ is given by Eq. (3.7) according to the corresponding experiments (see Sec. 3.2 and Ref. [4]).

Secondly, the change due to dispersal can be calculated by defining the dispersal kernel $\phi(\Delta_x, \Delta_y)$ as the probability per unit area that a virus initially placed at $(x+\Delta_x, y+\Delta_y)$ has moved to (x,y) after a time interval τ. Thus,

$$\begin{aligned}|[V](x,y,t+\tau)-[V](x,y,t)|_d \\ = \int\int [V](x+\Delta_x, y+\Delta_y, t)\phi(\Delta_x,\Delta_y)\,\mathrm{d}\Delta_x\mathrm{d}\Delta_y - [V](x,y,t).\end{aligned} \qquad (3.26)$$

In a system involving both reactive and dispersal processes, we add up their contributions

$$\begin{aligned}[V](x,y,t+\tau)-[V](x,y,t) \\ = \int\int [V](x+\Delta_x, y+\Delta_y, t)\phi(\Delta_x,\Delta_y)\,\mathrm{d}\Delta_x\mathrm{d}\Delta_y - [V](x,y,t) \\ + \tau F(x,y,t) + \left.\frac{\tau^2}{2}\frac{\partial F(x,y,t)}{\partial t}\right|_r + \cdots\end{aligned} \qquad (3.27)$$

Assuming that the kernel is isotropic, i.e., $\phi(\Delta_x, \Delta_y) = \phi(\Delta)$, with $\Delta = \sqrt{\Delta_x^2 + \Delta_y^2}$,and Taylor-expanding Eq. (3.27) up to second order in time and space,

$$\frac{\partial[V]}{\partial t} + \frac{\tau}{2}\frac{\partial^2[V]}{\partial t^2} = D\left(\frac{\partial^2[V]}{\partial x^2} + \frac{\partial^2[V]}{\partial y^2}\right) + F + \frac{\tau}{2}\frac{\partial F}{\partial t}\bigg|_r, \tag{3.28}$$

where $D = \frac{\langle\Delta^2\rangle}{4\tau} = \frac{\langle\Delta_x^2\rangle}{2\tau} = \frac{\langle\Delta_y^2\rangle}{2\tau}$ is the diffusion coefficient.

For large distances $r = \sqrt{x^2 + y^2}$ from the inoculation point of viruses $(x, y) = (0,0)$, $\frac{\partial^2[V]}{\partial x^2} + \frac{\partial^2[V]}{\partial y^2} \approx \frac{\partial^2[V]}{\partial r^2}$ and Eq. (3.28) is the same as Eq. (3.5), with F given by Eq. (3.7) and D replaced by D_{eff} (the reason for the latter change is explained in Sec. 3.3). Thus the terms proportional to τ in Eq. (3.5) arise simply from a second-order Taylor expansion. If the role of the eclipse time is neglected (τ), Eq. (3.28) reduces to the non-delayed or classical model used by Yin and co-workers [14, 144], namely [see Eq. (3.13)]

$$\frac{\partial[V]}{\partial t} = D\left(\frac{\partial^2[V]}{\partial x^2} + \frac{\partial^2[V]}{\partial y^2}\right) + F. \tag{3.29}$$

In general, adding up the reactive and diffusive contributions [as done in Eq. (3.27)] may not be exact [3, 145, 139, 183, 188] and this point is taken into account by the so-called sequential or cohabitation models (see especially Ref. [139], Fig. 1 of Ref. [183] and Fig. 17 of Ref. [188]). However, for virus infections cohabitation models yield almost the same results as non-cohabitation (or additive) models [145]. Thus in the present paper, we do not take the cohabitation effect into account for mathematical simplicity (the predicted speeds in Fig. 3.2 would be the same, so there is no need to use more complicated equations). Let us mention that, in contrast to virus infections, for human waves of advance the cohabitation effect is not negligible (and a more important effect still is due to the shape of dispersal kernels) [95, 183]. Such more precise models lead to the ballistic speed for fast reproduction [3, 179], as they should [3, 179, 139]. However, for virus infections those corrections are not necessary. In conclusion, the reaction-diffusion Eq. (3.5) has the microscopic derivation above and recent criticisms [139] are irrelevant. For τ and $F = 0$, this also provides a valid derivation of Fickian diffusion (Eq. (3.29) for $F = 0$). It is very important to stress that mathematical arguments [139] are not enough to establish whether a given equation is valid or not, because this depends on the system considered, and must thus be checked by using reactive functions, parameter values and initial conditions appropriate to the experimental setup (for example, to describe the growth of virus plaques, a model with only a pure death process is not realistic, and therefore irrelevant). As another example of this, reaction-diffusion with Fickian diffusion [Eq. (3.29)] can be applied if the delay time is negligible, which may be justified for some biological species but not for viruses. This is clearly seen in Fig. 3.2, by comparing our model to the classical or non-delayed one (3.12)-(3.14) used by Yin et al. [14, 144], which is based on Eq. (3.29). At the other extreme, the second-order approximation would obviously fail for large τ [4, 195, 139] and, if this happened, additional terms in the Taylor expansions above would be necessary. However, this is not our case (Sec. 3.10). Finally, Fickian diffusion [Eq. (3.29) with $F = 0$] can be applied if the diffusive delay time is sufficiently small and is useful in many situations (not in our case). Thus parameter values must be examined to choose the appropriate

equation for each experiment. Mathematical arguments are not enough, because an equation may be useful to describe some experiments but not others.

For completeness, in Sec. 3.10 we extend the derivation above to infinite order and find that the results are similar to those above and in the main paper (second order).

3.10. Full time-delayed equation

As shown in Sec. 3.9, Eq. (3.5) is in fact an approximation, because it includes only terms up to second order from the Taylor expansions. The virus density $[V]$ rapidly changes on a scale of time smaller than $\tau \approx 15$ min (because the increase in Fig. 3.1 take 6 minutes or less). This could therefore lead to errors in the front speeds obtained in Secs. 3.4 and 3.5. In this subsection we prove that this is not a problem by considering the full time-delayed equation [see Ref. [195], Eqs. (16) and (21)],

$$\sum_{n=1}^{\infty} \frac{\tau^n}{n!} \frac{\partial^n [V](r,t)}{\partial t^n} = \sum_{n=1}^{\infty} \frac{\left(2D_{eff}\tau\right)^n}{(2n)!} \frac{\partial^{2n} [V](r,t)}{\partial r^{2n}} + \sum_{n=1}^{\infty} \frac{\tau^n}{n!} \frac{\partial^{n-1} F(r,t)}{\partial t^{n-1}}\Bigg|_r, \tag{3.30}$$

instead of its approximation, Eq. (3.5), together with Eq. (3.6) and our new Eq. (3.4). Then, repeating the same steps as in Sec. 3.2 we get the following characteristic equation (which replaces Eq. (3.8)),

$$\left(e^{\lambda \bar{c}\bar{\tau}} - \cosh\left(\lambda\sqrt{2\tau}\right) - e^{-\kappa\bar{\tau}} + 1\right)\left(\lambda\bar{c} + e^{-\lambda\bar{c}\bar{\tau}}\right) = \frac{\kappa Y}{\lambda\bar{c} + \kappa}\left(1 - e^{-\kappa\bar{\tau} - \lambda\bar{c}\bar{\tau}}\right). \tag{3.31}$$

Repeating the calculations leading to Fig. 3.2, but using Eq. (3.31), we obtain that the differences are very small. Indeed, the error between the second-order approximation and full time-delay equation is lower than 3% for the three strains of the T7 virus. Thus, the use of the second-order approximation in Secs. 3.2-3.5 is valid.

4. A mathematical approach to virus therapy of glioblastomas[3]

Background It is widely believed that the treatment of glioblastomas (GBM) could benefit from oncolytic virus therapy. Clinical research has shown that Vesicular Stomatitis Virus (VSV) has strong oncolytic properties. In addition, mathematical models of virus treatment of tumors have been developed in recent years. Some experiments *in vitro* and *in vivo* have been done and shown promising results, but have been never compared quantitatively with mathematical models. We use *in vitro* data of this virus applied to glioblastoma.

Results We describe three increasingly realistic mathematical models for the VSV-GBM *in vitro* experiment with progressive incorporation of time-delay effects. For the virus dynamics, we obtain results consistent with the *in vitro* experimental speed data only when applying the more complex and comprehensive model, with time-delay effects both in the reactive and diffusive terms. The tumor speed is described by a very simple equation that nonetheless yields results within the experimental measured range.

Conclusions We have improved a previous model with new ideas and carefully incorporated concepts from experimental results. We have shown that the delay time τ is the crucial parameter in this kind of models. We have demonstrated that our new model can satisfactorily predict the front speed for the lytic action of oncolytic VSV on glioblastoma observed *in vitro*. We provide a basis that can be applied in the near future to realistically simulate *in vivo* virus treatments of several cancers.

Keywords biophysics, front propagation, mathematical model

4.1. Background

Since early last century, viruses have been studied as experimental agents for cancer treatment. The medical interest in the field has fluctuated during this period, reaching a fever pitch in the past two decades. It was in the early 1990s, when recombinant DNA technology became standard, that virus engineering could provide scientific furtherance to virotherapy. Then, oncolytic viruses appeared to be a treatment of tremendous potential and scientists started manipulating them to target cancerous cells more specifically. This culminated in the first marketing approval of an oncolytic virus, granted by the Chinese government in November of 2005 [29]. Very recently, improvements in patient survival have led to endorsements of other oncolytic virus in Europe and the US [30]. In parallel, mathematical

[3] This Chapter is an exact transcription of the contents of the following paper (please find a copy of the published version in Appendix B): de Rioja VL, Isern N, Fort J. A mathematical approach to virus therapy of glioblastomas. *Biology Direct* **11** 1-12 (2016). DOI: 10.1186/s13062-015-0100-7.

models of virus treatment of tumors have been developed [45, 126, 127]. However, even with this new ability to engineer viral genomes, a realistic therapeutic frontrunner has yet to emerge.

4.1.1. **Experimental background**

Among a variety of aggressive and deadly brain tumors we could highlight the glioblastoma. GBM is the most common and malignant brain cancer. Usually, treatment relies on chemotherapy, radiation and surgery. However these treatments are ineffective and the median survival time of a patient is no longer than 15 months (4 to 5 months without health care), due to multifocality of the disease, infiltrative growth and substantial tumor genotypic variability, among other factors [34, 23]. So, nowadays there are no known medical or surgical approaches that constitute an effective treatment of GBM, and for this reason it is widely considered that the treatment of GBM is likely to benefit from oncolytic virus therapy.

Oncolytic viruses –including retroviruses, herpesviruses and adenoviruses– are an emerging therapy tool for cancers that currently lack effective treatment [26]. The efficiency of different viruses against various tumor cell lines has been studied with *in vitro* and *in vivo* experiments [24, 196]. Of these, Vesicular Stomatitis Virus (VSV) has been shown in laboratory studies to have excellent capabilities to become one of the most valuable candidates for virotherapy, due to its very fast lytic cycle and its rapid oncolytic action. In addition, VSV is an enveloped, negative-strand RNA rhabdovirus that can infect a wide variety of species including mice and humans, though it is usually asymptomatic for human beings [25]. Therefore, the anticancer activity of mouse models can be transferable to human trials [118]. This fact makes VSV a strong oncolytic candidate and it has been used in preclinical studies of numerous cancer types, like glioblastomas.

Hence, we focus our attention on the development of a mathematical model of the VSV-GBM virus-tumor system. In the absence of *in vivo* data, all of the parameter values that we will introduce in the model are extracted from *in vitro* VSV-GBM experiments. Our main objective is to develop a simple model that can reproduce the VSV-GBM dynamics and explain satisfactorily the experimental *in vitro* propagation speeds.

4.1.2. **Previous mathematical approaches**

The most basic mathematical model of the competition between populations was constructed by Alfred J. Lotka and Vito Volterra in 1925 and 1926 independently [197]. For years their model was improved and adapted to different parasite-host systems, including virus infections [13, 14, 21, 159]. Nevertheless, we are interested in a specific model which studies the dynamics of an oncolytic virus through a tumor cell population.

In Ref. [45], Wodarz et al. noted that the few previous reaction-diffusion models of oncolytic virus spread [42, 43] include, in addition to basic spatial dynamics, one or more additional assumptions that introduce further complexity. In contrast, they opt for a very simple approach to the infection process with spatial dynamics. The process of adsorption of a virus V by a susceptible tumoral cell T (with rate k_1), and replication of Y viruses that leave each infected cell I (with rate k_2), is essentially described by the reactions

$$V + T \overset{k_1}{\rightarrow} I \overset{k_2}{\rightarrow} Y \cdot V. \tag{4.1}$$

Wodarz et al. study the behavior of an *in vitro* adenovirus in human embryonic kidney epithelial cells, experimentally and computationally, developing a simple model with two equations (see Eqs. (4.5) and (4.6) below), one for susceptible tumoral cells and one for infected cells. They make use of partial differential equations (PDEs) to model the virus-tumor system, because PDEs provide efficient information on the spatial and reactive mechanisms affecting the wave propagating fronts and PDEs can be used to compute their speeds.

The model by Wodarz et al. [45] is a two-equation system that was derived from a three-equation model due to Nowak and May [35]. Including diffusion and logistic growth, the Nowak-May model is

$$\frac{\partial [V](r,t)}{\partial t} = D_V \frac{\partial^2 [V](r,t)}{\partial r^2} + k_2 Y [I](r,t) - k_3 [V](r,t), \tag{4.2}$$

$$\frac{\partial [T](r,t)}{\partial t} = D_T \frac{\partial^2 [T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[I](r,t) + [T](r,t)}{k}\right\} - k_1 [V](r,t)[T](r,t), \tag{4.3}$$

$$\frac{\partial [I](r,t)}{\partial t} = D_I \frac{\partial^2 [I](r,t)}{\partial r^2} - k_2 [I](r,t) + k_1 [V](r,t)[T](r,t), \tag{4.4}$$

where $[T]$, $[I]$ and $[V]$ are the concentrations of susceptible tumoral cells, infected tumoral cells and viruses, respectively; D_T, D_I and D_V are their diffusion coefficients, a the tumor growth rate, k its carrying capacity, k_3 the decay rate of free viruses, t the time and r the radial coordinate (assuming radial symmetry, as explained in detail below). Some authors [35] have argued that, in some situations, it may be assumed that $\frac{\partial [V]}{\partial t} = 0$ and therefore, in homogeneous systems $\left(\frac{\partial^2 [V]}{\partial r^2} = 0\right)$, Eq. (4.2) implies that $[V](r,t) = \frac{k_2 Y}{k_3}[I](r,t)$. However, this assumption (free virus in steady-state) could only be applied if the decay rate of virus k_3 is much larger than the decay rate of the infected cell population k_2 [35]. From these arguments, they obtain the two-equation system used by Wodarz et al. [45], namely

$$\frac{\partial [T](r,t)}{\partial t} = D_T \frac{\partial^2 [T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[I](r,t) + [T](r,t)}{k}\right\} - b[I](r,t)[T](r,t), \tag{4.5}$$

$$\frac{\partial [I](r,t)}{\partial t} = D_I \frac{\partial^2 [I](r,t)}{\partial r^2} - k_2 [I](r,t) + b[I](r,t)[T](r,t), \tag{4.6}$$

where $b = \frac{k_1 k_2 Y}{k_3}$.

However, we find two drawbacks in the model (4.5)-(4.6) to explain our VSV-GBM system. First, Wodarz assumes $\frac{\partial [V]}{\partial t} = 0$, and thus $[V] \propto [I]$. As said before, this may be valid when $k_3 \gg k_2$ and in some non-spatial models [35] but this is in general not valid for the spatial propagation of virus infections. In such cases, at points located far away from the initially infected area, before the arrival of the infection front we have $[V] = 0$, when the infection arrives $[V] \neq 0$, and after all viruses (and infected cells) have decayed, we have again $[V] = 0$. Therefore, when dealing with spatial infection

fronts we have $\frac{\partial[V]}{\partial t} = 0$ only at early and late times, but $\frac{\partial[V]}{\partial t} > 0$ when the first viruses arrive and $\frac{\partial[V]}{\partial t} < 0$ after the passage of the infected front. Moreover, our experimental data (see Sec. 4.3) suggest that in our system k_3 is very close to k_2 and therefore, the assumption $k_3 \gg k_2$ is not satisfied here either. Therefore, in contrast to Ref. [45], we cannot assume $\frac{\partial[V]}{\partial t} = 0$, thus we deal with three differential equations (for viruses, susceptible tumoral cells, and infected tumoral cells).

Our second objection to the model (4.2)-(4.4) [and its simplification (4.5)-(4.6)] is that, according to the first reaction in Eq. (4.1), virus adsorption causes not only the same decrease in susceptible tumor cells [last term in Eq. (4.3)] as the increase in infected cells [last term in Eq. (4.4)], but also the same decrease in viruses. Thus a term $-k_1[V](r,t)[T](r,t)$ is missing in the right side of Eq. (4.2), in agreement with many previous works on virus infections [14, 21, 159, 130].

In the next section we develop a model which takes both points into account, as well as other important effects (namely, time-delay effects).

4.2. Methods

4.2.1. **Mathematical models**

Here we want to develop a simple, but complete model to understand the dynamics of a virus-tumor system. The theoretical model should be able to explain an *in vitro* experiment where a virus injected into the center of a tumor spreads through the tumor cell population in a basically two-dimensional geometry. Therefore, we can think of the virus-tumor system as formed by two fronts of propagation, which could be represented as two concentric circles if we assume radial symmetry. The diagram in Fig. 4.1 illustrates this idea. The outer circle represents the tumor cells, which spread to the outside through a non-specific medium. The inner circle represents the viruses spreading within the tumor. Viruses diffuse through the medium before infecting tumor cells. When infected cells die, a new generation of viruses is created and the process begins anew.

The main idea and experimental laboratory data come from Ref. [24], where Wollmann et al. compare nine types of viruses with strong oncolytic potential and conclude that four of them, VSV included, would be worthy of more rigorous studies. Because in subsequent papers [25, 155] they worked with VSV and its recombinant variants or strains, we decided to focus solely on VSV and use these data as experimental basis.

Below we present three increasingly complete (and complicated) models.

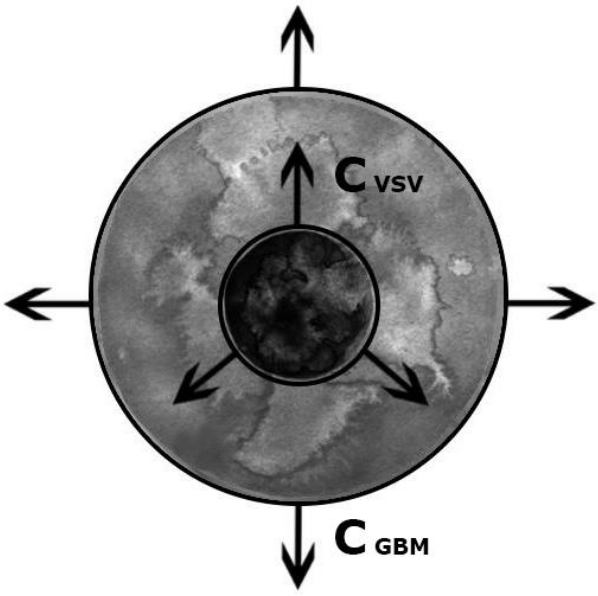


Figure 4.1 Two circles representing the two propagation fronts of VSV and GBM. A front of tumor cells spreads radially (large circle). After some time, viruses are inoculated at the center, and a virus front spreads (inner circle). If the inner circle expands faster than the outer one ($\boldsymbol{c_{VSV} > c_{GBM}}$), the viruses will eliminate the tumor.

Model 1

As a first approach, we adapt the model by Wodarz et al. [45] to the conditions in our VSV-GBM systems, i.e., we do not assume $\frac{dV}{dt} = 0$, and therefore $[V]$ is not proportional to $[I]$ and we need to include the virus dynamics explicitly in the model.

Now the evolution of the virus-tumor system is described by

$$\frac{\partial [V](r,t)}{\partial t} = D_{VSV} \frac{\partial^2 [V](r,t)}{\partial r^2} + F(r,t), \tag{4.7}$$

$$\frac{\partial [T](r,t)}{\partial t} = D_{GBM} \frac{\partial^2 [T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[I](r,t) + [T](r,t)}{k}\right\} - k_1[V](r,t)[T](r,t), \tag{4.8}$$

$$\frac{\partial [I](r,t)}{\partial t} = k_1[V](r,t)[T](r,t) - k_2[I](r,t). \tag{4.9}$$

The first equation describes the evolution of the virus population over time. The viruses can spread ruled by the diffusion coefficient D_{VSV} and the Laplacian (or second space derivative). The function $F(r,t)$ in Eq. (4.7) incorporates all processes of infection, replication and death and is defined by

$$F(r,t) = -k_1[V](r,t)[T](r,t) + k_2 Y[I](r,t) - k_3[V](r,t). \tag{4.10}$$

Note that the first term was not included in the models by Nowak-May and Wodarz [Eq. (4.2)] (see our second objection in Sec. 4.1.2).

Eq. (4.8) describes the change in the number of tumor cells over time. Similarly to viruses, glioblastoma cells can also move, characterized by their own diffusion coefficient D_{GBM}.

Finally, Eq. (4.9) represents the evolution of infected tumoral cells. We assume that these cells do not move, in agreement Fig. 3D of Ref. [24], where the experiment shows how the infected cells (U-87 MG glioblastoma cells) initially introduced do not move through the host layer throughout the observation period.

Model 2

As we shall see in Sec. 4.4, model 1 needs further improvements. In model 2 we take into account that infected tumoral cells do not die instantaneously, instead there is a time delay before the cell dies and releases the new progeny of viruses. We will denote this delay or eclipse time as τ and include it into the terms related to the death of infected cells. Thus infected cells will not die proportionally to the density of infected cells at the present time, $k_2[I](r,t)$, but proportionally to the density of infected cells at a previous instant $t-\tau$, $k_2[I](r,t-\tau)$, to properly include this time delay effect on the decay process. It has been shown that the term $-k_2[I](r,t-\tau)$ agrees well with experimental data in a different context (infections of non-tumor cells) [7]. Other reaction-diffusion models do also apply $t-\tau$, although in an alternative way [147, 146]. The differences between their approach and ours are analyzed in Ref. [7].

Therefore, when introducing the delay in the death of infected cells, Eqs. (4.9) and (4.10) are modified directly and Eq. (4.7) changes because the function $F(r,t)$, Eq. (4.14), is also modified. We do not modify the growth term in Eq. (4.8) because the reproduction of tumoral cells depends on the total number of tumor cells (infected and susceptible) at that precise instant t. So, we consider the model

$$\frac{\partial[V](r,t)}{\partial t} = D_{VSV}\frac{\partial^2[V](r,t)}{\partial r^2} + F(r,t), \tag{4.11}$$

$$\frac{\partial[T](r,t)}{\partial t} = D_{GBM}\frac{\partial^2[T](r,t)}{\partial r^2} + a[T](r,t)\left\{1-\frac{[I](r,t)+[T](r,t)}{k}\right\} - k_1[V](r,t)[T](r,t), \tag{4.12}$$

$$\frac{\partial[I](r,t)}{\partial t} = k_1[V](r,t)[T](r,t) - k_2[I](r,t-\tau), \tag{4.13}$$

where now

$$F(r,t) = -k_1[V](r,t)[T](r,t) + k_2 Y[I](r,t-\tau) - k_3[V](r,t). \tag{4.14}$$

This second model is, actually, an approximation of our next model (see model 3 below).

Model 3

Model 2 takes into account a delay time in the reactive process $I \to Y \cdot V$, but here we shall see that the delay time also has a very important diffusive effect. The diffusion dynamics of the virus concentration in Eq. (4.11) is Fickian, which means that it does not take into account the effect of the time delay τ. In year 2002 it was shown [4] that it is very important to take into account that τ is the time interval during which a virus does not move in space (because it is inside an infected cell), thus the delay time should affect the model by slowing down the spread of viruses. Therefore it is necessary

to incorporate also this effect to reach a realistic model. For this reason, Eq. (4.11) must be replaced by an equation with second-order terms to include this diffusive time-delay effect [4, 21, 117].

Thus, finally we describe the spatial-time dynamics of the whole system with the following equations:

$$\frac{\partial[V](r,t)}{\partial t}+\frac{\tau}{2}\frac{\partial^2[V](r,t)}{\partial t^2}=D_{VSV}\frac{\partial^2[V](r,t)}{\partial r^2}+F(r,t)+\frac{\tau}{2}\frac{\partial F(r,t)}{\partial t}\bigg|_g, \tag{4.15}$$

$$\frac{\partial[T](r,t)}{\partial t}=D_{GBM}\frac{\partial^2[T](r,t)}{\partial r^2}+a[T](r,t)\left\{1-\frac{[I](r,t)+[T](r,t)}{k}\right\}-k_1[V](r,t)[T](r,t), \tag{4.16}$$

$$\frac{\partial[I](r,t)}{\partial t}=k_1[V](r,t)[T](r,t)-k_2[I](r,t-\tau), \tag{4.17}$$

where the terms proportional to τ in Eq. (4.15) are the new, second-order terms. A self-contained derivation of Eq. (4.15) can be found in Ref. [7], Appendix A.

In Eq. (4.15) $F(r,t)$ is again given by Eq. (4.14), and Eqs. (4.12) and (4.13) from model 2 remain unchanged [Eqs. (4.16) and (4.17), respectively].

Note that $F(r,t)$ can be understood as the variation of $[V]$ over time due to all reactive processes, but not to diffusive processes, i.e. $F(r,t)=\frac{\partial[V](r,t)}{\partial t}\Big|_g$. This allows the proper calculation of the first time derivative as [21, 117]

$$\frac{\partial F(r,t)}{\partial t}\bigg|_g=-k_1F(r,t)[T](r,t)-k_1[V](r,t)\frac{\partial[T](r,t)}{\partial t}+k_2Y\frac{\partial[I](r,t-\tau)}{\partial t}-k_3F(r,t). \tag{4.18}$$

For systems in which the infected cells diffuse appreciably (not our case, see the last paragraph in the model 1 section), an age-structure model with this additional diffusive-delay effect has been proposed by Gourley and Kuang in Ref. [146], p. 558.

In the equation describing the virus dynamics, Eq. (4.15), we include corrections only up to second order [21, 117]. It has been shown in previous work [4] that the divergence between second-order approximation and full time-delayed equations is small, and thus we can exclude terms of higher orders.

4.2.2. **Front speeds**

Virus front

Using models 1-3 above, we look for realistic travelling-wave speeds for both the propagation front of viruses (inner front, Fig. 4.1) and the propagation front of tumor cells (outer front, Fig. 4.1). Finding the propagation speeds will allow us to compare to the *in vitro* experiments in order to validate our approach.

In all models 1-3, we can transform the problem into a single-variable system by using the co-moving coordinate $z=r-ct$. Like in previous works [4, 14], we assume the concentration of the three

populations at the leading edge of the moving front ($z \to \infty$) can be written as $[T] = k - \epsilon_T \cdot e^{-\lambda z}$, $[I] = \epsilon_I \cdot e^{-\lambda z}$ and $[V] = \epsilon_V \cdot e^{-\lambda z}$, thus we assume that tumoral cells are nearly at maximum concentration at large distances from the inoculation point of the viruses, while viruses and infected cells are barely present. We make use of this transformation because beyond the edge of the front of infected cells and viruses, there is only a continuous medium of tumor cells. For non-trivial solutions to exist, the determinant of the matrix corresponding to the linearized model must be zero. The characteristic equations for model 1, model 2 and model 3 are, respectively,

$$(\lambda c + k_2)(\lambda c - D_{VSV}\lambda^2 + kk_1 + k_3) - kk_1k_2Y = 0, \tag{4.19}$$

$$\left(\lambda c + k_2 e^{-\lambda c\tau}\right)(\lambda c - D_{VSV}\lambda^2 + kk_1 + k_3) - kk_1k_2Ye^{-\lambda c\tau} = 0, \tag{4.20}$$

$$\begin{aligned}\left(\lambda c + k_2 e^{-\lambda c\tau}\right)\left[\lambda c - D_{VSV}\lambda^2 + kk_1 + k_3 + \frac{\tau}{2}(\lambda^2 c^2 - k^2k_1^2 - 2kk_1k_3 - k_3^2)\right] \\ - kk_1k_2Ye^{-\lambda c\tau}\left[1 + \frac{\tau}{2}(\lambda c - kk_1 - k_3)\right] = 0,\end{aligned} \tag{4.21}$$

According to marginal stability analysis [189], the propagation front moves with the minimum possible speed. Therefore,

$$c_{VSV} = min_{\lambda>0}[c(\lambda)], \tag{4.22}$$

where $c(\lambda)$ is given implicitly by Eqs. (4.19), (4.20) and (4.21). From Eq. (4.22) we can numerically estimate the speed of VSV infection.

The resulting propagation speeds for models 1-3 will be calculated and plotted in Sec. 4.4.

We also solve the third model by numerical integration and find the front speed from the position of the virus front wave in successive steps of time.

Glioblastoma front

Under the hypothesis of two propagation fronts, as shown in Fig. 4.1, the outermost front would corresponds the tumor cells, $[T]$ (GBM in our case of study). In the conditions near this front, all models can be greatly simplified since here the populations of viruses and infected cells are zero (see the outer circle in Fig. 4.1 for a better understanding), so $[V](r,t) = 0$ and $[I](r,t) = 0$. Hence, it is only necessary to work with the equation for the tumoral cells, Eq. (4.16) for example, but remembering that $[V](r,t) = [I](r,t) = 0$,

$$\frac{\partial[T](r,t)}{\partial t} = D_{GBM}\frac{\partial^2[T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[T](r,t)}{k}\right\}. \tag{4.23}$$

At the leading edge of this front, we assume that $[T](r,t) = \epsilon_T \cdot e^{-\lambda z}$, and after some algebra we easily obtain the speed of the glioblastoma front,

$$c_{GBM} = 2\sqrt{D_{GBM}a}, \tag{4.24}$$

where D_{GBM} is the glioblastoma diffusion coefficient and a the growth rate, both estimated in the next subsection. Note that Eq. (4.24) is the well-known Fisher propagation speed [96]. Some recent

extensions have been proposed [34, 198], but they are not necessary for the purposes of the present paper.

4.3. Parameter values

We estimate most of our parameters from *in vitro* experiments on VSV applied to GBM [25, 24, 155]. The parameters that we could not draw from such experiments have been obtained from other rigorous studies on VSV or glioblastoma.

We use two different values of D_{VSV} because the diffusion coefficient of VSV has not been measured in gliomas. The only value of VSV available (measured in an specific water solution) is $D_{VSV} = 8.37 \cdot 10^{-5}$ cm²/h [151]. Another value measured in agar of VSV-similar viruses is $D_{VSV} = 1.44 \cdot 10^{-4}$ cm²/h [21].

Concerning D_{GBM}, Stein et al. [152] performed an *in vitro* experiment in which a GBM tumor spheroid is implanted into a collagen gel. The diffusion coefficient is measured by tracking individual cells on the first day, calculating their motion and averaging over many cells. Stein and co-workers measure diffusion coefficients in the radial and angular directions, which lead to the value $D_{GBM} = 3.75 \cdot 10^{-6}$ cm²/h [34].

Besides spreading, the number of cells also increases. The parameter a is the corresponding proliferation rate. *In vitro* measurements provide ample scope for this parameter, $0.04 < a < 0.3$ day^{-1} [153], and similarly *in vivo* studies yield $0.01 < a < 0.14$ day^{-1} [161].

The saturation cell density, k, measures the maximum concentration of tumor cells (susceptible and infected) per unit volume that the system can support, and its usual value is $k = 10^6$ cells/cm^3 (e.g., Refs. [163, 162]).

We next analyze the rest of parameters, which are calculated from the experimental studies by Wollmann et al. [25, 24, 155].

The yield or burst size Y represents the total amount of viruses produced by the death of a single infected cell. There is no accurate numerical value calculated for the case of VSV infecting GBM. However, by studying Fig. 4 in Ref. [25] we can obtain an estimation. The burst size can be understood as the ratio between the maximum and initial number of viruses, i.e. $Y = \frac{V_{max}}{V_0}$. From that figure, V_0 is between 10-100 PFU/ml (last two plots in Fig. 4 in [25]) and V_{max} between 10^8 – 10^9 PFU/ml (the maximum is reached between 1 and 2 days post-infection), so we conclude that $10^6 < Y < 10^8$. This also agrees with the value measured in Ref. [158], although in that case VSV infects BHK-21 cells (not GBM cells).

We have seen that there is a time lapse between a cell being infected by a virus and that cell dying (and therefore, adding more viruses to the system). This time lapse is called the delay time, τ. It plays a main role in the virus propagation speed, but has not been accurately measured. From the *in vitro* experiments described in Ref. [24] we can try to estimate the value of τ. On one hand, we know that the death of infected cells begins about 6 hours post-infection (hpi) of the virus to susceptible tumoral cells. We also know that infected cells can be seen as early as 4 hpi (they are tracked down using GFP fluorescence). From both data, we conclude that viruses leave infected cells at least 2 h after infection.

On the other hand, in a different experiment infected cells are added directly (rather than infecting viruses) and new infected cells were detected after 12 h. This period includes the time needed for the viruses to multiply within the infected cells, leave the cell and infect new tumoral cells. So we can also assume that τ must be lower than 12 h. In summary, we will work with the range $2 < \tau < 12$ h.

The adsorption rate, k_1, describes the efficacy of the whole infection process (rate of virus entry and probability of successful infection). The value of k_1 could be measured in an experiment where the reproduction of viruses and host cells were prevented. Such an experiment has been performed for other viruses [154] but not for VSV infecting GBM. Since we do not have the ideal conditions in the experiments cited before [25, 24, 155], we will use the earliest data post-inoculation available in the experimental data in Ref. [25] to minimize the effect of reproduction and thus obtain the best possible estimation for k_1.

Eqs. (4.7) and (4.8) are simplified in the absence of reproduction and natural death, and when the population is studied as a whole (i.e. ignoring diffusion terms) we have

$$\frac{d[V](t)}{dt} = \frac{d[T](t)}{dt} = -k_1[V](t)[T](t). \tag{4.25}$$

Obviously, integrating we get $[T](t) = [V](t) + \xi$, where ξ is the constant of integration. Note that ξ is the difference between the concentrations of tumor cells and viruses. In order to estimate k_1, we can rewrite the previous Eq. (4.25) as $\frac{d[T](t)}{dt} = -k_1[T](t)([T](t) - \xi)$ and making the necessary algebra we obtain the final formula for calculating the adsorption rate,

$$k_1 = \frac{1}{\xi(t - t_0)}\left[\ln\left(\frac{T}{T - \xi}\right) - \ln\left(\frac{T_0}{T_0 - \xi}\right)\right]. \tag{4.26}$$

It is difficult to know the exact concentration of cells at the beginning of the experiment or at certain time t, because only relative concentrations were reported. However, extrapolating data provided in the previous cited papers by Wollmann et al. (Fig. 3C Control in [25], bar G/GFP), we believe it is correct to assume that the values of initial tumor cells lie in the range $T_0 = 10^6 - 10^8$ cells/cm^3, and that $T = 0.65T_0$ cells/cm^3, $t - t_0 = 36$ h. This allows the calculation of the adsorption rate, as $5 \cdot 10^{-10} < k_1 < 5 \cdot 10^{-8}$ cm^3/h. This is a rather wide range, but we show in Sec. 4.4.2 that k_1 (as well as Y) does not overly affect the propagation front speed of VSV.

Finally, parameters k_2 and k_3 correspond to the rates of death of infected cells and virus, respectively. Therefore, the average life-time of an infected cell and a virus are $1/k_2$ and $1/k_3$, respectively.

The rate of death of infected cells k_2 could be also understood as the growth of viruses. Thus, for $t < \tau$ no new virus are seen in the corresponding experiment (because no infected cell has died yet), but for $t \geq \tau$ the infected cells start to die ruled by $dI = -k_2 I_0 dt$. The death of each infected cell produces Y virus, thus $dV = -YdI = k_2 Y I_0 dt = k_2 V_{max} dt$. Integrating, we get $k_2 = \frac{V_{max} - V_0}{\Delta t \cdot V_{max}} \approx \frac{1}{\Delta t} = \frac{1}{t^* - \tau}$, where t^* represents the time when the virus population reaches its maximum. According to Fig. 4B in Ref. [25], experimental data (labeled as VSV-G/GFP) show that the maximum is reached at $t^* = 48 \pm 12$ h. Nevertheless, the final result of k_2 will depend on τ and we have a range rather than a single value for τ (see above). Note, however, that for model 1 there is no time delay, so k_2 is

calculated straightforwardly as the inverse of time t^* at which the concentration of viruses reaches its maximum, $k_2 = \frac{1}{t^*}$ h^{-1}. Models 2 and 3 are dealt with in Sec. 4.4.

The evolution of the viruses over time in an environment where they die but cannot reproduce is ruled by $dV = -k_3 V dt$. Through simple integration we get $V(t) = V_0 e^{-k_3(t-t_0)}$. In the same experiment as before, Fig. 4B in Ref. [25], we now have two cases where these conditions are exactly reproduced (because VSV-dG-GFP and VSV-dG-RFP are replication-restricted virus variants, so they basically die). We can estimate both values of k_3 from the experimental data, namely $V(t = 24\text{h}) = 30$ PFU/cm^3, $V(t = 48\text{h}) = 20$ PFU/cm^3 and $V(t = 72\text{h}) = 8$ PFU/cm^3 for the mutant dG-GFP and $V(t = 24\text{h}) = 12$ PFU/cm^3, $V(t = 48\text{h}) = 8$ PFU/cm^3 and $V(t = 72\text{h}) = 6$ PFU/cm^3 for dG-RFP. Performing linear fits to $\ln V$ versus t, we obtain that $0.014 < k_3 < 0.028$ h^{-1}.

4.4. Results and discussion

4.4.1. GBM and VSV front speeds: theory versus experiment

Our main objective is to obtain realistic values for the propagation speeds in an *in vitro* virus-tumor system, providing positive results from a biophysical point of view for the realization of these treatments.

In Sec. 4.2 we have described three possible models for our VSV-GBM system and the necessary experimental parameter values. Here we present the speeds predicted by these models.

The case of tumor expansion has a single, simple solution for all models, Eq. (4.24), since the infection does not play a role here. Substituting the values of D_{GBM} and a we obtain that $c_{GBM} = 2.5 \cdot 10^{-4}$ cm/h, with $a = 0.1$ day^{-1}, which we think is a reasonable mean value. Indeed, the range of measurements of the proliferation rate is $0.01 < a < 0.3$ day^{-1}, which yields a range of speeds $7.9 \cdot 10^{-5} < c_{GBM} < 4.33 \cdot 10^{-4}$ cm/h). Stein and co-workers measured an experimental *in vitro* speed range of $2.37 \cdot 10^{-4} < c_{GBM} < 5.54 \cdot 10^{-4}$ cm/h [153], which is consistent with our model, despite the simplicity of Eq. (4.24).

The case of the virus front is less straightforward. As we have already discussed in Sec. 4.3, a very important but not strictly well-measured parameter is the delay time τ. Therefore, the speed results have been calculated in terms of this parameter, $c(\tau)$. The death rate of infected cells k_2 also changes, because it depends directly on τ.

The infection front speed, c_{VSV}, can be seen in Fig. 4.2. For each of the 3 models we have plotted the results from typical parameter values (bold lines). To compute these results we have chosen the parameter values that seem to be the most representative and accepted for this experiment: average values of k_2 and k_3, the value of D_{VSV} calculated for VSV in an specific water solution and the larger values of k_1 and Y. However we have also computed c_{VSV} by varying each of the parameters of Eqs. (4.19)-(4.21), with the exception of k because $k = 10^6$ cells/cm^3 is a widely accepted value in research papers (see Sec. 4.3). In Fig. 4.2 we include the upper and lower bounds for the front speed obtained, for each of the 3 models, from the experimental parameter ranges (parameter values are specified at the caption).

The hatched area in Fig. 4.2 corresponds to the experimental values of VSV speed estimated from the *in vitro* experiment by Wollmann et al. in Ref. [24], Fig. 3A.

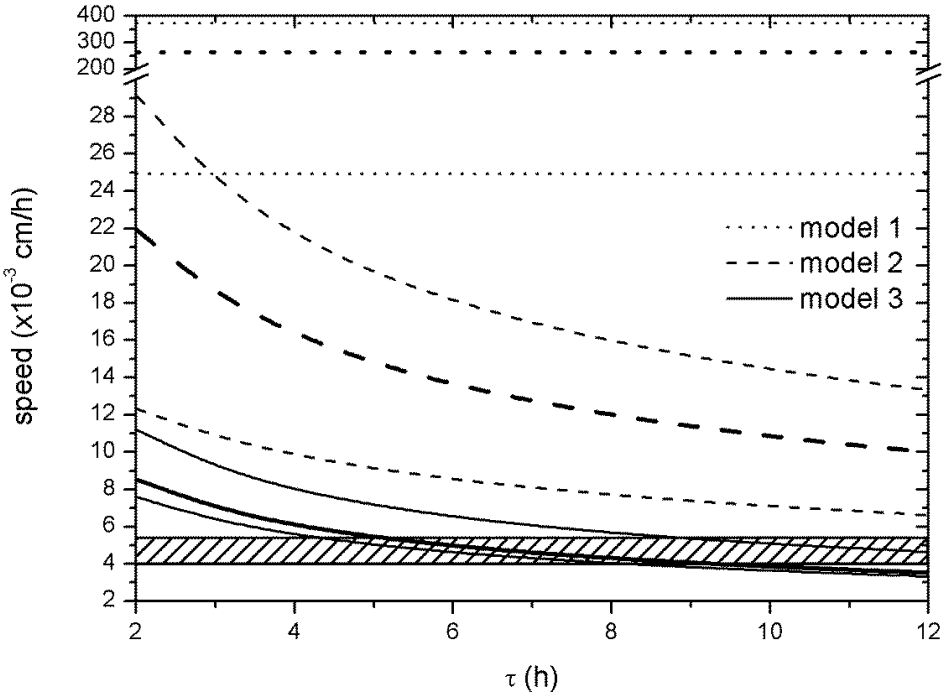


Figure 4.2 VSV front propagation speed as a function of the delay time $\boldsymbol{\tau}$, for model 1 (dotted lines), model 2 (dashed curves) and model 3 (solid curves). The hatched area shows the experimental *in vitro* VSV front speed [24]. Upper bounds are computed for: $\boldsymbol{k_1 = 5 \cdot 10^{-8}}$ cm^3/h, $\boldsymbol{k_2 = \frac{1}{36-\tau}}$ h^{-1} ($\boldsymbol{k_2 = \frac{1}{36}}$ h^{-1} for model 1), $\boldsymbol{k_3 = 0.014}$ h^{-1}, $\boldsymbol{Y = 10^8}$ and $\boldsymbol{D_{VSV} = 1.44 \cdot 10^{-4}}$ cm^2/h. Lower bounds are computed from: $\boldsymbol{k_1 = 5 \cdot 10^{-10}}$ cm^3/h, $\boldsymbol{k_2 = \frac{1}{60-\tau}}$ h^{-1} ($\boldsymbol{k_2 = \frac{1}{60}}$ h^{-1} for model 1), $\boldsymbol{k_3 = 0.028}$ h^{-1}, $\boldsymbol{Y = 10^6}$ and $\boldsymbol{D_{VSV} = 8.37 \cdot 10^{-5}}$ cm^2/h. The results from typical values (bold lines) are computed from: $\boldsymbol{k_1 = 5 \cdot 10^{-8}}$ cm^3/h, $\boldsymbol{k_2 = \frac{1}{48-\tau}}$ h^{-1} ($\boldsymbol{k_2 = \frac{1}{48}}$ h^{-1} for model 1), $\boldsymbol{k_3 = 0.02}$ h^{-1}, $\boldsymbol{Y = 10^8}$ and $\boldsymbol{D_{VSV} = 8.37 \cdot 10^{-5}}$ cm^2/h. In all the cases $\boldsymbol{k = 10^6}$ cells/cm^3.

Dotted lines correspond to the analytical results to model 1, Eqs. (4.7)-(4.10), i.e. the classical model adapted from the equations in Ref. [45]. Obviously they are horizontal lines, since they do not depend on τ. As we can see in Fig. 4.2, model 1 yields speeds much faster than the experimental observations. The curves are the numerical results from our time-delayed reaction-diffusion models. Dashed curves correspond to model 2, given by Eqs. (4.11)-(4.14). We see that just by taking into account the eclipse or delay time on the death of infected cells, we obtain much better results as compared with experimental velocities, although not enough to satisfactorily explain the data (the minimum bound of model 2 in Fig. 4.2 is above the hatched area). Finally, solid curves in Fig. 4.2 correspond to model 3 (please recall that this is extremely close to the full time-delayed equation, see Sec. 4.2). The equations for this main model, Eqs. (4.15)-(4.18), when considering typical parameter values, produce results that agree with the experimental data within a range of τ between 5.0 h and 9.3 h.

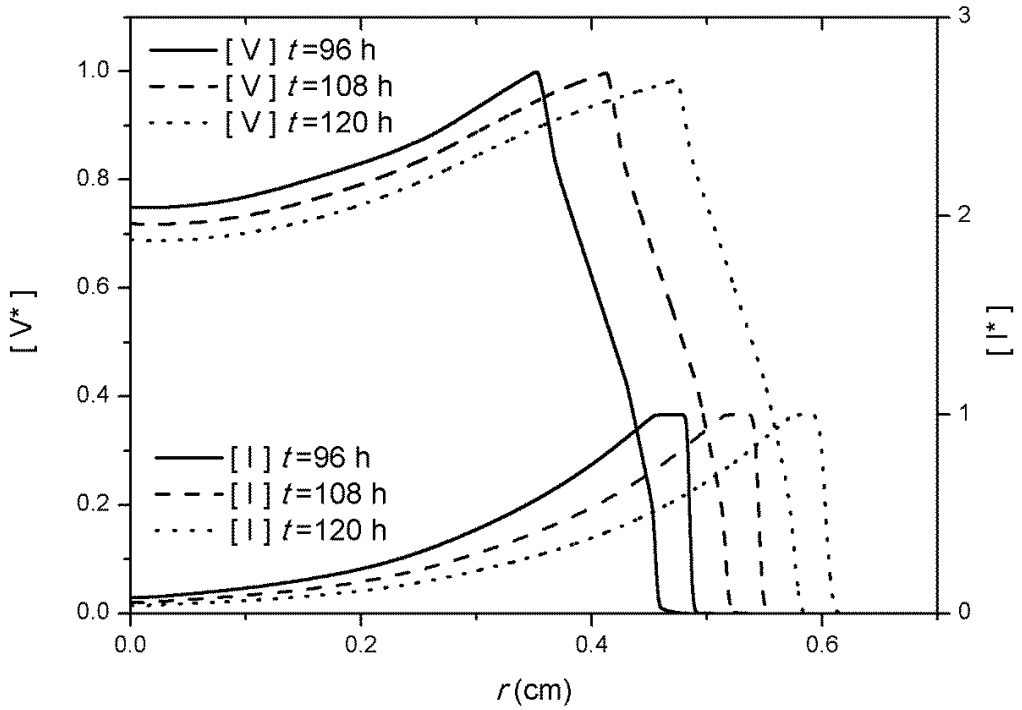


Figure 4.3 Radial profiles of $[\boldsymbol{V}^*]$ and $[\boldsymbol{I}^*]$ at three different times for model 3. The labels of $\boldsymbol{V}^*$ and $\boldsymbol{I}^*$ stand for the units used, defined as $\frac{[\boldsymbol{V}]}{[\boldsymbol{V}]_{max}}$ and $\frac{[\boldsymbol{I}]}{[\boldsymbol{I}]_{max}}$, respectively. The profiles are computed from numerical integration.

According to our best description (model 3), the entire range of speed c_{VSV} in Fig. 4.2 is an order of magnitude faster than the speed of propagation of glioblastoma c_{GBM} (see above). Therefore the virus front could theoretically reach the tumoral front and infect it all. We stress that this is a model appropriate for *in vitro* experiments, whereas *in vivo* more complex models will be necessary (as discussed below).

In Fig. 4.3 we show snapshots of the viruses and infected cells profiles as functions of the radial axis, computed from the computational simulations at three time instants. The simulations have been performed by numerical integration of model 3, which is biologically more realistic and produces results in agreement with the experimental data (see Fig. 4.2). We use the typical parameter values used in Fig. 4.2 (bold lines, see caption for the values). We can see in Fig. 4.3 that both propagation fronts advance at the same speed and with regular shapes.

From the profiles we can see that the number of infected cells grows rapidly, then there is a plateau of infected cells (as a result of the time delay τ before any infected cell dies), and then decay at a rate k_2. The virus profiles show an abrupt rise when infected cells start dying (end of the plateau of infected cells) and then keep rising up to a peak. Behind this peak, the virus death term k_3 predominates over the virus production, and the number of viruses decays. Although Fig. 4.3 seems to indicate that the front of infected cells appears prior to the virus front, the opposite happens (this can be appreciated by enlarging the vertical scale).

From these simulations we can calculate the front speed by tracking the position of the edge of the front of the virus at successive steps of time. A simple space vs time data is generated and then, the front speed is directly the slope. From the simulations (parameter values are the same than typical

values in Fig. 4.2 with $\tau = 6$ h) we find a front speed of $4.829 \cdot 10^{-3}$ cm/h. The relative error between the simulations and the analytic speed [$c_{VSV} = 4.853 \cdot 10^{-3}$ cm/h, from Eqs. (4.21) and (4.22)] is only about 0.5%.

An alternative way to know the front propagation speed from Fig. 4.3 is the plateau of infected cells. Its width is directly related with the time delay τ and the infection front speed as $width = \tau \cdot c$. Then, the result for the speed is $(0.53858 - 0.51317)\text{cm}/6\text{ h} = 4.735 \cdot 10^{-3}$cm/h (distances for $t = 108$ h), and the relative error (compared with the analytical results with same parameter values than the simulations) is less than 2.5% ($c_{VSV} = 4.853 \cdot 10^{-3}$ cm/h).

4.4.2. Effects of k_1 and Y

In Sec. 4.3 we have estimated the values of the parameters used in our mathematical models. Some of them, e.g. D_{VSV}, D_{GBM} and k, have well-defined values, which are taken from the references indicated in the text. The delay time τ plays a very important role and therefore we have found the front propagation speed as a function of this parameter (remember that $k_2 = \frac{1}{48-\tau}$, so we could add k_2 to this argument). Other parameters like a and k_3 have a range of possible values, albeit a narrow one, and as such we use the mean value, or that usually accepted by other sources. Lastly, parameters Y and k_1 have very wide ranges, spanning several orders of magnitude, but as we shall show below, they do not have an important effect on the virus front speed.

In Fig. 4.4 the speed of VSV is calculated from model 2 [Eqs. (4.11)-(4.14)] and model 3 [Eqs. (4.15)-(4.18)]. Setting the typical parameter values previously used in Fig. 4.2 (bold curves) and Fig. 4.3 for D_{VSV}, D_{GBM}, k, k_3 and the average value $\tau = 8$ h (so $k_2 = \frac{1}{40}$ h^{-1}), which yields results consistent with the range of experimental speeds (Fig. 4.2), we have varied the values of Y and k_1 for each of both models.

In model 2 (upper curves in Fig. 4.4) the speed dependence on Y and k_1 is fairly important. Indeed, by increasing these variables by two orders of magnitude, the speed increases on average by 25% and 18%, respectively. However, looking at the best approach, model 3 (lower curves), we note that the speed increases only by 3% and 2% for Y and k_1, respectively.

Therefore, model 3 has little dependence on the parameters Y and k_1 and the delay time is the most important parameter (Fig. 4.2). In contrast, model 2 depends more directly on both parameters, although τ still remains the crucial one (compare the change of the speed in Fig. 4.2 with those in Fig. 4.4 for model 2). To obtain a speed of virus propagation similar to the observed data ($c \approx 5 \cdot 10^{-12}$ cm/h) with model 2, we should modify Y and k_1 out of the experimental ranges. Indeed, their values should be about $Y = 10^4$ or $k_1 = 5 \cdot 10^{-12}$ cm^3/h. Therefore, we could get a speed in agreement with the experimental data, but only using unrealistic parameter values, which do not correspond to VSV. This is further proof that our final model 3, the time-delayed reaction-diffusion set of equations, is a good mathematical tool to explain this kind of virus-tumor biological systems.

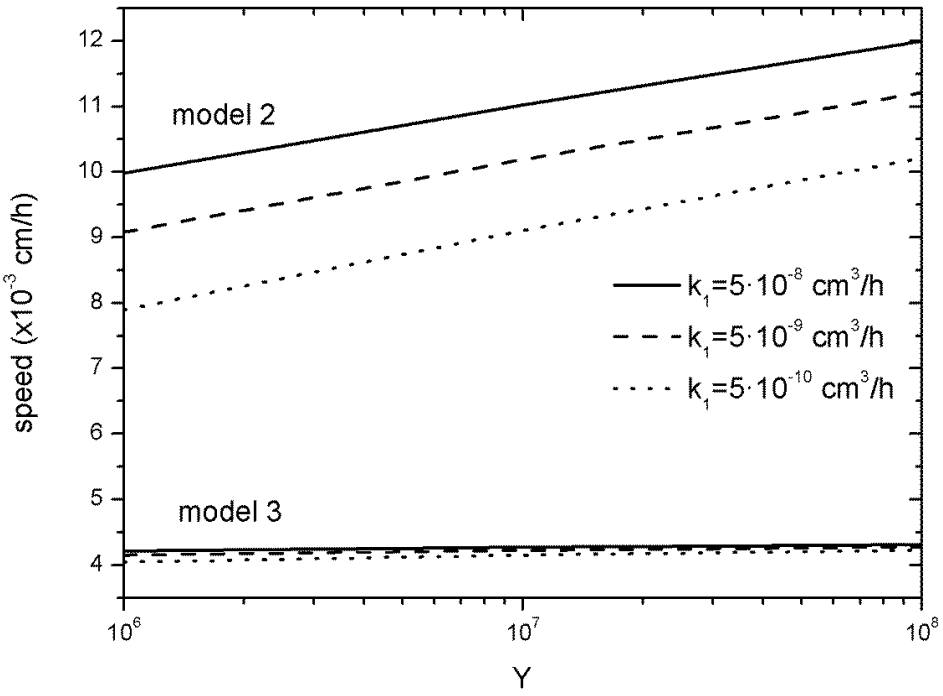


Figure 4.4 VSV invasion speed on GBM for various values of $\boldsymbol{Y}$ and $\boldsymbol{k_1}$. The other parameter values are $\boldsymbol{k} = \boldsymbol{10^6}$ cm^{-3}, $\boldsymbol{k_2} = \frac{\boldsymbol{1}}{\boldsymbol{40}}$ h^{-1}, $\boldsymbol{\tau = 8}$ h and $\boldsymbol{k_3 = 0.02}$ h^{-1}. Model 3 proves that neither $\boldsymbol{Y}$ nor $\boldsymbol{k_1}$ affect much the speed of the front.

4.5. Conclusions

A simple set of time-delayed equations have been built to understand the dynamics of a virus-tumor system. We have improved a previous model with new ideas and carefully incorporated experimental results (especially Ref. [24]). Figure 4.2 proofs that our best framework (model 3) is in reasonable agreement with the experimental data. Furthermore, the figure shows that neither model 1 nor model 2 can explain the experimental data. So it is absolutely necessary to add the second-order terms proportional to τ in Eq. (4.15) to properly include the time-delay effect.

We have shown that the delay time τ is the crucial parameter in our models (even when compared to other parameters that are strongly unknown, such as k_1 and Y). As we could have expected, as τ increases, the speed of the virus front decreases, because viruses spend more time inside the cell, and therefore at rest. In spite of being of utmost importance, the role of the delay or eclipse time has not been taken into account in previous models of virus treatment of tumors [45, 42, 43].

We have found that our new model can satisfactorily predict the front speed for the lytic action of oncolytic VSV on glioblastoma observed *in vitro*. But this is only a first step towards a deep biophysical understanding of the principles of virus-tumor space-time spread in a complex system. This model could be extended to be applied to *in vivo* experiments where, among other effects, the immune response should be also included in the model because it may play a significant role regulating the efficacy of the therapy. In particular, it seems that there is currently no agreement about which approach is better in oncolytic therapy, whether to modify oncolytic viruses to obtain the maximum antitumoral immune response [199], to transiently suppress the immune response [118], or to use a

combination of both [118]; future appropriate modeling of the three scenarios might help in tackling this controversy from a different perspective.

In this paper we have focused on GBMs because of the experimental data available, but our model could apply also to many non-diffusive cancers, for which viral therapy is a promising approach [42, 43, 200], since the reaction-diffusion equations for the viruses [Eqs.(4.15)-(4.18)] will still be valid, even though in such cases tumor cells will not diffuse. Thus, we provide a basis that can be applied in the near future to realistically simulate *in vivo* virus treatments of several cancers.

4.6. Acknowledgments

This work was partially funded by ICREA (Academia award) and the MINECO (projects SimulPast-CSD2010-00034, FIS-2009-13050 and FIS-2012-31307).

5. The ancient cline of haplogroup K implies that the Neolithic transition in Europe was mainly demic[4]

Abstract Using a database with the mitochondrial DNA (mtDNA) of 513 Neolithic individuals, we quantify the space-time variation of the frequency of haplogroup K, previously proposed as a relevant Neolithic marker. We compare these data to simulations, based on a mathematical model in which a Neolithic population spreads from Syria to Anatolia and Europe, possibly interbreeding with Mesolithic individuals (who lack haplogroup K) and/or teaching farming to them. Both the data and the simulations show that the percentage of haplogroup K (%K) decreases with increasing distance from Syria and that, in each region, the %K tends to decrease with increasing time after the arrival of farming. Both the model and the data display a local minimum of the genetic cline, and for the same Neolithic regional culture (Sweden). Comparing the observed ancient cline of haplogroup K to the simulation results reveals that about 98% of farmers were not involved in interbreeding neither acculturation (cultural diffusion). Therefore, cultural diffusion involved only a tiny fraction (about 2%) of farmers and, in this sense, the most relevant process in the spread of the Neolithic in Europe was demic diffusion (i.e., the dispersal of farmers), as opposed to cultural diffusion (i.e., the incorporation of hunter-gatherers).

Keywords genetic clines, Neolithic transition, Europe, ancient DNA, forager-farmer interaction

5.1. Introduction

The Neolithic transition was a major transformation that introduced agricultural economics, radically changed the environment, and led to increased population densities and new forms of social organization [49, 201]. The Neolithic spread from the Near East across Europe, from about 8,000 yr Before the Common Era (BCE) until about 3,000 yr BCE [47]. A crucial question is whether the spread of the Neolithic was due to a dispersal of farming populations (demic diffusion), to the learning of agricultural techniques by European hunter-gatherers (cultural diffusion), or to a combination of both mechanisms. The latter possibility is suggested by the comparison of archaeological data to mathematical wave-of-advance models, which indicate that demic diffusion was more important than cultural diffusion [3, 95]. Genome-wide studies also indicate a crucial role for demic diffusion, with very little cultural diffusion at the beginning of the Neolithic [80, 81]. Notwithstanding the unquestionable importance of genome-wide studies, it is also of interest to analyze specific genetic markers, for two reasons. First, genome-wide studies cannot provide any quantitative explanation for the observed spatial cline of a single marker. And secondly, genome-wide studies cannot yield a quantitative estimate for the percentage of farmers involved in cultural diffusion. In order to understand both limitations of genome-wide studies, consider first one marker that has not been

[4] This Chapter is an exact transcription of the contents of the following paper (please find a copy of the published version in Appendix B): Isern N, Fort J, de Rioja VL. The ancient cline of haplogroup K implies that the Neolithic transition in Europe was mainly demic, *Scientific Reports* **7** 11229, 1-10 (2017).

affected by drift neither selection. If there is admixture between the populations of incoming farmers and indigenous hunter-gatherers (HGs), and the latter originally lacked this marker, then it will dilute progressively, i.e. its frequency will decrease with increasing distance from the spatial region of origin of the Neolithic front. Second, consider again a maker unaffected by drift neither selection, but such that HGs initially had higher frequencies than farmers. Its frequency will not decrease but increase with distance from the Neolithic origin. Thirdly, consider a marker that increased its frequency after some location during the spread of the Neolithic front (due, e.g., to surfing or other drift effects). If HGs originally lacked this marker, its frequency will decrease (due to admixture) up to some distance and increase for larger distances. Fourthly, if several drift and/or selective effects took place, the cline can have even more complicated shapes. Thus, clearly the frequencies of different genetic markers have different spatiotemporal dependencies, because they are due to different processes. For this reason, in order to estimate the percentage of farmers involved in cultural diffusion we should not to include many arbitrary markers (as in genome-wide studies). Instead, we should consider very specific markers that satisfy the following conditions: (1) the frequency decreases with increasing distance from the spatial origin of the Neolithic front; (2) HGs lack the marker considered before the arrival of the first farmers (otherwise we would need to know the precise space-time variation of the marker initial frequency in HGs); (3) selection and (4) drift (including surfing) effects can be neglected. This makes it possible to compare the data to demic-diffusion models neglecting drift, selection, etc. (as done below). In the present paper we analyze mitochondrial haplogroup K because, as we shall see, the observed data for this marker satisfy conditions (1) and (2). In contrast, other markers that have been found in Early Neolithic European sites (e.g., N1a, J, T and X) have not been found in the Near East [165], so condition (1) does not hold. Condition (3) can be also reasonably assumed, because there are no data indicating the existence of any selective pressure on haplogroup K, and analysis of the Early Neolithic K haplotypes does not show signs of selection (Sec. 5.8.1). It is also reasonable to assume that condition (4) holds, because we will show that a simulated cline (neglecting drift) is consistent with the observed one for haplogroup K.

A totally independent reason why genome-wide studies cannot determine quantitatively the percentage of farmers involved in cultural diffusion is that, e.g., Mathieson et al. [80] assume only two source populations, Anatolian Neolithic and Western HG, and use f_4-statistics to estimate, e.g., a 93% of Anatolian Neolithic ancestry and a 7% of Western HG ancestry for Early Neolithic farmers in Germany. But this result of 93% is not the percentage of farmers involved in cultural diffusion. Instead, it is the Anatolian fraction ($\alpha_1 = 0.93$) of genetic drift (defined as a variance of allele frequencies [202]) of the German population considered (assuming that its drift is a linear combination of the drifts of the two presumed source populations). But there is no mathematical theory relating the proportions of genetic drift (i.e., the coefficients $\alpha_1, \alpha_2, \ldots \alpha_N$ of the f_4-value of a test population in terms of the f_4-values of its N presumed source populations [80, 202]) to the percentage of farmers involved in cultural diffusion. Similarly, in admixture analysis the fractions of the genome contributed by a set of presumed source populations are estimated, but again there is no theory relating them to the percentage of farmers involved in cultural diffusion. For totally analogous reasons, these and other previous methods (f_4-statistics, admixture, principal components, structure analysis, *D*-statistics, etc.) can provide valuable qualitative indications on whether demic or cultural diffusion dominated the Neolithic spread, but they cannot yield any quantitative value for the percentage of farmers involved in cultural diffusion. Incidentally, we note that many such methods (e.g., f_4-statistics and admixture) assume a few source populations, whereas here we will consider the more realistic case of populations distributed continuously in space (possibly with seas and mountains). If clinal patterns are not observed in analyses based on principal components, admixture, f_3, f_4, *D*-statistics, etc. (where, instead, early Neolithic individuals tend to cluster together, e.g. with modern Sardinians), the reason is simply that those analyses are based on many markers but, as explained above, the spatial

distribution of each one can be due to other processes in addition cultural diffusion (surfing, other kinds of drift, selection, etc.).

In this article we shall estimate the percentage of early farmers involved in cultural diffusion from an ancient DNA (aDNA) marker. We will perform our analysis at the continental scale, because aDNA data are not yet numerous enough to consider specific geographic regions. As we shall see, however, there are already sufficient data to obtain some first estimates of general trends. We consider mitochondrial DNA (mtDNA), because nuclear data are known for a substantially smaller number of ancient individuals. Mitochondrial DNA is inherited from the mother, thus its study will be related to the spread of maternal lineages. As all genetic sequences, mtDNA can be inherited with mutations, but similar sequences (haplotypes) with a common ancestor are usually grouped into haplogroups. Since the aDNA data are still limited in number, we perform our analysis below at the haplogroup level, grouping together the different haplotypes and subclades from each lineage (in Sec. 5.8.1 we include analyses at the haplotype level, and they reinforce our conclusions). The mtDNA of European hunter-gatherers is composed mainly of U lineages (U, U4, U5, and U8), which are absent in early Neolithic populations [78, 203, 204]. Conversely, haplogroups N1a, T2, K, J, HV, V, W, and X have been proposed as potential Neolithic markers because they have been found in farmers of the Linearbandkeramic (LBK) culture, an early Neolithic culture in Central Europe, and are almost absent in neighboring hunter-gatherer samples [78, 169]. Haplogroup K has been identified in only three hunter-gatherers dated before the arrival of farming (two in Greece [171] and one in Georgia [204]), but their subclades have not been found so far in any Neolithic farmer (see Sec. 5.8.2 for a detailed discussion of the very few exceptions of Mesolithic individuals displaying K haplotypes). Thus haplogroup K was virtually absent in Europe before the arrival of farming, and condition (2) above is satisfied. On the other hand, as we shall see below, haplogroup K displays a cline of decreasing frequency with increasing distance from the spatial origin of the Neolithic expansion. Thus haplogroup K also satisfies condition (1) above, in contrast to other potential Neolithic markers (N1a, T2, J, HV, V, W, and X).

5.2. Results and discussion

In order to study the existence of a genetic cline for haplogroup K in early Neolithic populations and subsequently compare it to our simulations, we have gathered all mtDNA information of Early and Middle Neolithic individuals reported in the literature, and we have grouped the data into regional cultures according to their location, date and reported culture (Appendix A Data S1). The Neolithic expansion in Europe begun in the Near East, and for this reason we have used the oldest pre-pottery Neolithic B (PPNB) date from Syria [47], Ras Shamra, as a geographic reference for the origin of the spread. In Fig. 5.1 we represent, for each regional culture, the average date of its individuals whose mtDNA haplogroup has been determined against the distance from their average location to Ras Shamra. Figure 5.1 includes all regional cultures dated between the Early and the Middle Neolithic, such that the mtDNA haplogroup of more than two individuals is known (e.g., Greece could not be included; see Sec. 5.8.3). The Southern Levant is not included for reasons explained in Sec. 5.8.3. For each regional culture, the number of individuals whose mtDNA haplogroup has been determined is given in the caption to Fig. 5.1 (Appendix A Data S2 and Data S3). We distinguish 3 different groups of regional cultures in Fig. 5.1. The first group is composed by the 10 *oldest* Neolithic regional cultures (from Syria to western and northern Europe) for which there are genetic data (squares in Fig. 5.1). The second group (triangles in Fig. 5.1) corresponds to 15 regional cultures that have *younger* dates than the oldest ones (squares) and that are located at similar distances from Syria (i.e., broadly in the same area). Thus, the triangles in Fig. 5.1 are not representative of the earliest local Neolithic cultures. Finally, the circle in Fig. 5.1 corresponds to Sweden. Its date and location are those of the earliest Neolithic individuals in Sweden whose mtDNA is known. It would be thus legitimate to consider this data point (circle in Fig. 5.1) simply as one of the oldest regional cultures (squares), and we will actually

include it into our calculations below. But the date for Sweden is substantially delayed relative to other cultures located at similar distances (Fig. 5.1), so it will be useful to identify Sweden with a symbol (circle) different than the other oldest regional cultures (squares).

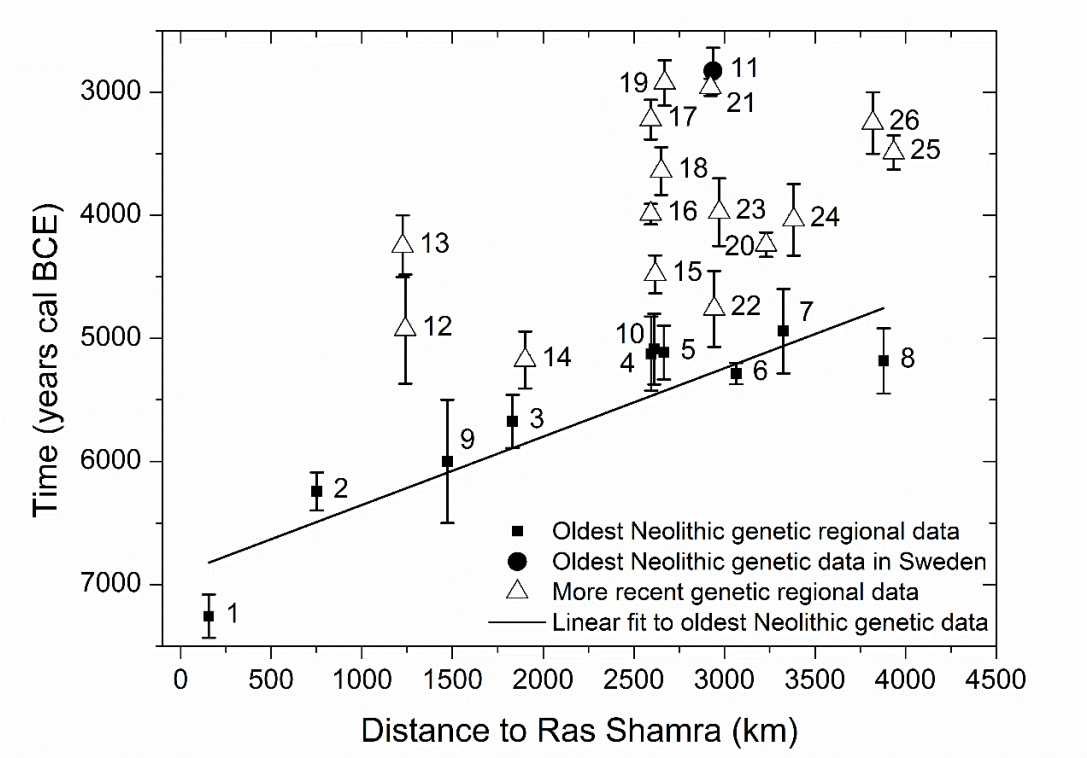


Figure 5.1 Dates versus great-circle distances from Ras Shamra (Syria) for 26 regional cultures with ancient mtDNA data. Squares correspond to the oldest regional Neolithic cultures, namely 1 Syria PPNB (15 individuals), 2 Anatolia (28 individuals), 3 Hungary-Croatia Starčevo (44 individuals), 4 Eastern Germany LBK (36 individuals), 5 Western Germany LBK (56 individuals), 6 Northeastern Spain Cardial (15 individuals), 7 Spain Navarre (36 individuals), 8 Portugal coastal Early Neolithic (10 individuals), 9 Romania Starčevo (5 individuals) and 10 Southern Germany LBK (4 individuals). The circle stands for 11 Sweden (9 individuals), which is substantially delayed due to the slowdown of the Neolithic front in northern Europe. Triangles correspond to more recent regional cultures, namely 12 Romania Middle Neolithic (29 individuals), 13 Romania Late-Middle Neolithic (9 individuals), 14 Hungary LBK (45 individuals), 15 Eastern Germany RSC (10 individuals), 16 Eastern Germany SCG/BAC (38 individuals), 17 Eastern Germany SMC (30 individuals), 18 Western Germany BAC (14 individuals), 19 Western Germany BEC (17 individuals), 20 Western France Prissé (3 individuals), 21 South-Eastern France Treilles (29 individuals), 22 Catalonia Epicardial (7 individuals), 23 Catalonia Late Epicardial (3 individuals), 24 Spain Basque country (7 individuals), 25 Portugal coastal Late Neolithic (3 individuals) and 26 Portugal inland Late Neolithic (7 individuals). The straight line is the regression fit to the 10 oldest regional data (squares).

5.2.1. Understanding the observed variations in the percentage of haplogroup K (%K)

It is important to keep in mind that the *oldest* regional cultures displayed in Fig. 5.1 do not correspond to the oldest archaeological dates known for each Neolithic regional culture, but only to the oldest Neolithic individuals whose mtDNA haplogroup has been determined. In spite of this, those dates (squares in Fig. 5.1) show a highly linear dependence on distance (correlation coefficient $R = 0.93$), as predicted for the oldest dates by wave-of-advance models [3]. In Fig. 5.2 we plot the %K as function of distance from Ras Shamra (Syria) for all the regional cultures in Fig. 5.1 that include at least 8 individuals (regions with fewer individuals have been ignored to avoid very large error bars). Because

the total number of individuals per region is still small in many regions, in our analysis below we take into account the whole 80% confidence-level (80% CL) range, represented as error bars, rather than only mean values. Labels and symbols in Fig. 5.2 are the same as in Fig. 5.1. For the oldest Neolithic cultures, there is no theoretical reason to expect a linear dependency of the %K on distance (in other words, we should not expect a high value of P for the regression to the squares and circle in Fig. 5.2). However, the slope of this regression in Fig. 5.2 is highly significantly different from zero ($P = 0.001$), and this gives statistical support to the existence of a genetic cline (similarly, low values of P also yield statistical support to the existence of phonemic clines [6, 205]). Additional support to the existence of a cline is obtained from an interpolation map and the analysis of the %K data by means of a Moran's I correlogram, included in Sec. 5.8.4.

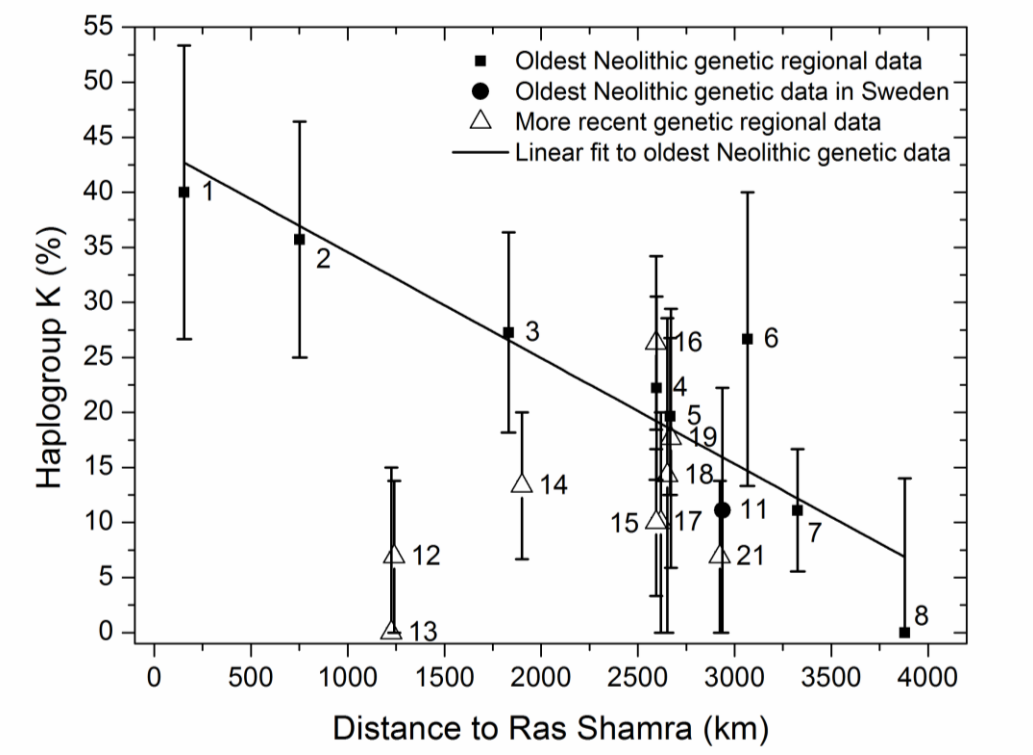


Figure 5.2 Observed percentage of mtDNA haplogroup K as a function of the great-circle distance from Ras Shamra (Syria). Each number denotes the same culture as in Fig. 5.1 (regions with fewer than 8 individuals have been ignored to avoid very large error bars). The straight line is the regression fit to the 10 oldest regional data (squares) and the oldest data in Sweden (circle). Error bars display 80% CL intervals (see Sec. 5.4.2).

As explained in the Introduction, it has been proposed that haplogroup K spread across Europe from the Near East with the Neolithic front [78, 165, 166, 172, 173]. We shall call this proposal the wave-of-advance model of haplogroup K. The analysis of the Early Neolithic haplotypes in our database also yields support to the assumption that the population with haplogroup K underwent a recent process of demographic and geographic expansion (Sec. 5.8.1). Obviously demic diffusion, on its own, cannot explain the spatiotemporal distribution of haplogroup K (as displayed in Fig. 5.2), because purely demic diffusion predicts a uniform distribution (see below). Thus we ask whether cultural (in addition to demic) diffusion is a viable explanation. If HGs lacked haplogroup K (as justified by genetic data in the Introduction and Sec. 5.8.2), and other effects (selection, drift, mutation, etc.) can be neglected, a demic-cultural model makes two testable predictions. (i) For the earliest Neolithic cultures, we should observe a decrease in the percentage of farmers with haplogroup K with increasing distance from the Near East (because interbreeding of pioneer farmers with local hunter-gatherers, and/or acculturation of the latter during the front propagation, will diminish the %K). This prediction is clearly observed in Fig. 5.2, because the earliest regional Neolithic cultures (squares and circle) show a clear

decrease of the %K with increasing distance from Syria. (ii) For each region, this model also predicts that the earliest Neolithic regional culture will have a higher percentage of farmers with haplogroup K than later cultures (due to interbreeding and acculturation subsequent to the arrival of the Neolithic wave of advance). Prediction (ii) is also observed in Fig. 5.2, because of the 9 European cultures that do not correspond to the earliest Neolithic (triangles), only 1 (culture 16) has a larger %K than the expected regional maximum (the latter is given by the linear fit to the earliest regional Neolithic cultures in Fig. 5.2), and even culture 16 may be lower than the expected maximum, if the error bar is taken into account. However, we must caution that prediction (ii) refers to populations dated substantially later than the spread of the Neolithic front and it is therefore affected by population movements and other processes subsequent to the spread of the Neolithic. Thus, it is not reasonable to try to explain *quantitatively* prediction (ii) with a simulation model of the spread of the Neolithic. For this reason, although the model satisfies *qualitatively* both predictions, in the rest of this paper we shall be mainly concerned with prediction (i).

5.2.2. **Ancient cline of haplogroup K**

Figure 5.3 shows (lines) the clines obtained from our wave-of-advance simulations of the Neolithic and haplogroup K spread (see Sec. 5.4 and Secs. 5.8.5-5.8.6), alongside the observed genetic data for the earliest regional Neolithic cultures (squares and circle) already depicted in Fig. 5.2. In Fig. 5.3 we have imposed the initial genetic conditions that all simulations predict the observed %K for Syria (square labelled 1; see more details on the implementation of the initial conditions in Sec. 5.4 and Sec. 5.8.7). The simulated clines have been computed at the same 9 locations and dates as the genetic data (so the lines simply join the 9 data points), and for several values of the cultural diffusion intensity η.

We first observe that, similarly to the behavior of the data (symbols in Fig. 5.3) and in agreement with prediction (i) formulated above, when considering cultural diffusion ($\eta \neq 0$), the %K from the simulations (lines) tends to decrease with increasing distance from the Near East. This behavior was to be expected, because more distance from the origin (Ras Shamra, Syria) implies more time for the farming populations to interact (via interbreeding or/and acculturation) with hunter-gatherers (who lack haplogroup K). However, we note that both the simulations and the data display a local minimum at region 11 (Sweden). This is due to the fact that, according to archaeology [47] and ancient genetics [167, 206], the spread of the Neolithic in Europe occurred following two main routes: one along the Mediterranean coast (corresponding to the Impressa and Cardial traditions) and the other through the Balkans and the Central European plains (corresponding to the Starčevo and LBK cultures). To see how this explains the minimum in Fig. 5.3, consider first the Neolithic front propagating along the Mediterranean coast. In this case, population dispersal is driven by jumps (maritime migrations) of about 150 km per generation (see Sec. 5.4 and Sec. 5.8.6; in agreement with previous simulation results [47]). Conversely, the Neolithic front propagating inland is driven by jumps of about 50 km per generation (Sec. 5.4). Therefore, in order for the Neolithic front to travel a given distance, a *coastal* propagation obviously implies fewer jumps, i.e., fewer generations, and therefore less time for interbreeding with hunter-gatherers (and/or acculturation of the latter) than an *inland* propagation. Thus a *coastal* route will lead, at a given distance, to a lower decrease of the %K than an *inland* route. This is why the Mediterranean route leads, in region 6 (NE Spain) in Fig. 5.3, to higher values of the %K than the central-northern European route in region 11 (Sweden), in spite of the fact that the former is further away from Syria than the latter. This explains the minimum in the simulation curves (and in the observed data) in Fig. 5.3 (see Sec. 5.8.8 for a more detailed discussion, and Fig. 5.18 for a plot of the simulated clines along both routes).

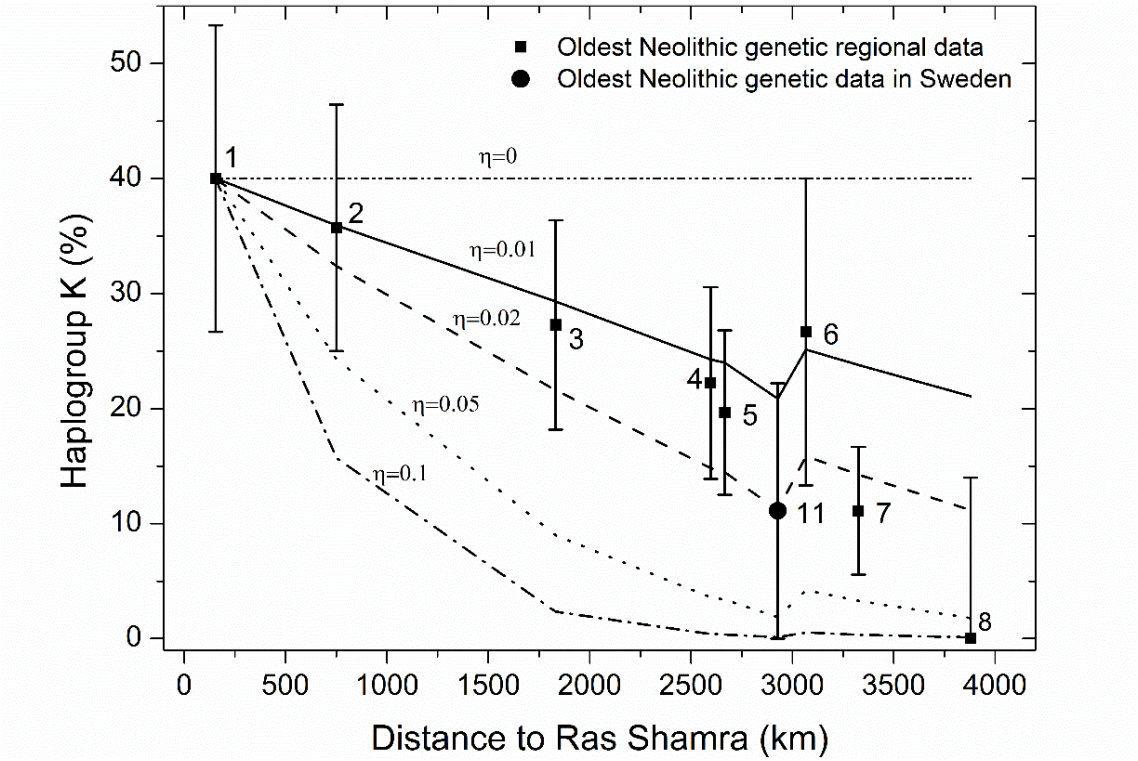


Figure 5.3 Observed and simulated percentage of mtDNA haplogroup K as a function of the great-circle distance from Syria. The data are shown with the same error bars as in Fig. 5.2, but only for the oldest regional cultures. The lines are the results of the mathematical simulation for several values of the cultural diffusion intensity η. The lines have been plotted by joining the simulation results for each of the 9 regional cultures, obtained at the average location and date of the individuals whose mtDNA haplogroup has been determined for each regional culture (Appendix A Data S1). Therefore, the simulation result for each region has been obtained at its average date (Fig. 5.1 and Appendix A Data S1). Numerical labels denote the same cultures as in Figs. 5.1-5.2.

5.2.3. **Demic versus cultural diffusion**

What does the observed cline of haplogroup K for Early Neolithic cultures (error bars in Fig. 5.3) imply about the importance of cultural diffusion in the spread of the Neolithic? First, let us examine how the intensity η of cultural diffusion is related to the steepness of the genetic cline. Note that, in the absence of cultural diffusion (i.e., without interbreeding neither acculturation), the %K at all farming populations would remain approximately constant at the value observed for the original (PPNB) population in Syria (assuming that drift and other processes do not have a strong effect). Thus, in a purely demic model ($\eta = 0$), such a cline would not be observed. Accordingly, the simulation for $\eta = 0$ leads to a uniform distribution in Fig. 5.3. We also expect that the stronger the intensity of cultural diffusion, the more important the decrease in the frequency of haplogroup K, and the steeper its geographic cline. This intuitive expectation agrees with the simulation results in Fig. 5.3, where for any given distance from the origin, a higher value of the cultural transmission intensity η yields a lower %K.

By comparing the data (symbols) to the demic-cultural space-time simulations (lines), we observe that Fig. 5.3 implies that the intensity of cultural diffusion was $\eta \approx 0.02$ (because higher or lower values of η lead to lines that are not within all of the error bars obtained from the aDNA data). The maximum possible value of this parameter is $\eta = 1$ [94] [see the text below Eq. (5.2)]. Therefore, although the observed cline cannot be explained without cultural diffusion ($\eta = 0$, horizontal line in Fig. 5.3), such a low value ($\eta \approx 0.02$) implies that cultural diffusion was remarkably weak. Indeed, the

cultural diffusion intensity η can be interpreted as the proportion of pioneering farmers that mate a hunter-gatherer [94] or, alternatively, that teach agriculture to a hunter-gatherer [3] (Sec. 5.8.9). Thus, our result that $\eta \approx 0.02$ (Fig. 5.3) implies that cultural diffusion involved only a tiny fraction (about 2%) of farmers and, in this sense, the most relevant process in the Neolithic spread in Europe was demic diffusion. Modifying the initial conditions so that the whole 80% CL for Syria is considered refines this estimate of the percentage of farmers involved in cultural transmission to the range $(2 \pm 1)\%$ (Sec. 5.8.7). The primacy of demic diffusion has been noted in genome-wide studies (see, e.g., previous work by Mathieson et al. [80]), but those studies could not quantify the percentage of farmers involved in cultural diffusion (see our Introduction). In contrast, we quantify that about 98% of farmers did not take part in cultural diffusion.

Our main result, namely that a very small amount of cultural transmission is enough to produce a continent-wide genetic cline, agrees with previous simulations [5, 88, 89], which however did not use the equations of cultural transmission theory nor could compare to aDNA data (which were then also unavailable). Therefore, in none of those previous studies was it possible to estimate quantitatively the percentage of farmers involved in cultural diffusion.

5.3. Conclusions

In this paper we have analyzed the genetic implications of a mathematical model that combines demic dispersal, population growth, and cultural transmission theory. Using anthropologically realistic assumptions and parameter values, we have performed, to the best of our knowledge, the first qualitative and quantitative comparison of a mathematical model to an observed Neolithic genetic cline. Although the ancient genetic data currently available are still limited, especially those corresponding to the Early Neolithic, they cover a wide enough area (see Sec. 5.8.4, Fig. 5.10) to allow us to analyze the geographical cline of genetic markers at the continental level, even if regional variations cannot be detected. In addition, the data are numerous enough so that we can observe a cline, and reach conclusions valid at least at the 80% CL (error bars in Fig. 5.3 and in Sec. 5.8.7, Fig. 5.15 and Fig. 5.17). A Moran's I correlogram confirms the existence of the cline (Sec. 5.8.4, Fig. 5.11). We have focused our attention on haplogroup K, mainly because it is virtually absent in hunter-gatherer populations and its frequency has a maximum in the Near East (specifically in Syria). Both points make it possible to attempt a description based on a simple mathematical model.

Qualitatively, the model predictions agree with the data in two ways: (i) both the data and the simulations show that the %K tends to decrease with increasing distance from Syria (Fig. 5.3); (ii) for each region, the %K tends to decrease with increasing time after the arrival of farming (Fig. 5.2).

Quantitatively, comparison between the model and the data shows that: (i) both the model and the data display a local minimum of the genetic cline, and for the same regional culture (Sweden, i.e. symbol 11 in Fig. 5.3); (ii) the ancient cline of haplogroup K can be explained if about 98% of farmers were not involved in cultural diffusion. However, we stress that the observed cline cannot be understood assuming that 100% of farmers were not involved in cultural diffusion. Thus, the observed cline implies that some farmers took part in cultural transmission (either by interbreeding or by teaching agriculture to hunter-gatherers). But only a tiny fraction (about 2%) of farmers were involved in cultural diffusion. In this sense, the most relevant process in the expansion of Neolithic culture in Europe was demic diffusion, i.e. the reproduction and dispersal of farmers, as opposed to the incorporation of hunter-gatherers (cultural diffusion).

Recently, the conclusion that the spread of the Neolithic in Europe was driven mainly by demic diffusion has been also obtained from comparing non-genetic, demic-cultural models to the spread rate of the Neolithic front, as estimated from archaeological data [3]. However, using only archaeological data has severe limitations. The reason is the following. Archaeological data make it

possible to estimate the spread rate of the Neolithic wave of advance, and this can be compared to the results of the mathematical model. But the dependence of the spread rate on the intensity of cultural transmission is weak [3, 94] and, for this reason, the spread rate can be used only to estimate an upper bound for the intensity of cultural transmission (namely $0 < C < 2.5$ [3], equivalent to $0 < \eta < 2.5$ here, see Sec. 5.8.9). In contrast, here we have shown that genetic data make it possible to know a function that depends strongly on the intensity η of cultural transmission (Fig. 5.3), namely the percentage of the considered haplogroup as a function of distance (i.e., the genetic cline shown in Fig. 5.3). This strong dependency has made possible a much more precise estimation of the percentage of farmers involved in cultural diffusion, namely $\eta = 0.02$ (Fig. 5.3), i.e. about 2%. This shows the tremendous potential of combining genetics, archaeology and mathematical modelling. On the other hand, the high number of archaeological data has allowed the identification of regional variations [95], something that is still not possible on the basis of ancient genetic data.

Our findings agree with genome-wide results, in the sense that demic diffusion was the main driver of the Neolithic spread in Europe (see, e.g. the results by Mathieson et al. [80]). However, genome-wide studies cannot estimate the percentage of farmers involved in cultural diffusion (see our Introduction). In contrast, our methodology yields the first quantitative estimation for this percentage (about 2%). This is possible because, in contrast to genome-wide studies, our approach has two crucial features: first, we compare to cultural-demic wave-of-advance mathematical models; second, we use a marker that shows decreasing frequency with increasing distance from the Near East. This estimate arises from comparing our model to the data at the 80% CL, leading to a confidence interval for the importance of cultural diffusion of $(2 \pm 1)\%$. Of course, if additional such markers are identified in future work, they will yield more precise results and will also allow the study of regional variabilities. Thus the present paper is a first step, which also provides a plausible explanation for the observed cline of haplogroup K at a continental scale. We stress that such an explanation cannot be provided by genome-wide studies. For simplicity, our models assume the same dispersal behavior for males and females. If future studies detect ancient clines of decreasing frequency for additional genetic markers, and they consistently show differences between maternal and paternal markers, they could be used to infer different dispersal behaviors for females and males, using trivial extensions of our models.

Ancient DNA data indicate that cultural diffusion was more important in some specific regions, such as Scandinavia [177] or the Paris Basin [207]. Thus, it has been recently suggested that the effect of cultural diffusion increased as farmers migrated farther west in Europe [207]. This suggestion agrees nicely with: (i) our simulated clines (lines in Fig. 5.3); (ii) the observed cline of haplogroup K (symbols in Fig. 5.3); and (iii) the intuitive expectation that longer distances from the spatial origin of the Neolithic imply more time for interbreeding and/or acculturation and, therefore, a stronger effect of cultural diffusion.

5.4. Materials and methods

5.4.1. Archaeological and genetic data

We gathered a database of all individuals from farming cultures dated between 8,000 and 3,000 calibrated years BCE for which the mtDNA haplogroup have been reported in the literature. For all 513 individuals in the database, we report the haplogroup, date, latitude, longitude, bibliographical references and additional data (Appendix A Data S1). We grouped them into regional cultures

according to their geographical and cultural closeness (e.g., Syria PPNB, Anatolia, Hungary-Croatia Starčevo, Hungary LBK, etc.). The data from Syria are from PPNB sites, which makes them especially relevant because PPNB/C are the Near-Eastern Neolithic cultures that later spread into Europe [47]. We selected for further analysis the 26 regional cultures with more than two individuals (comprising 508 individuals) and discarded the others (see Sec. 5.8.3 for a discussion on Neolithic individuals not included in the analysis). For each of the 26 selected regional cultures, we calculated the percentage of individuals with K haplotypes (Appendix A Data S2 and Data S3), the average date of its individuals, and the average great-circle distance of its individuals to the site of Ras Shamra (Appendix A Data S3). This is the oldest PPNB Syrian site used in previous simulations studies [47], and we therefore use it as origin of the Neolithic range expansion in our simulations (see below).

5.4.2. **Statistical analysis**

For each of the 26 regional cultures, we estimated the error intervals of its average date and %K. The time error bar (Fig. 5.1) was estimated as the range of dates for all individuals in the considered regional culture. The error bar for the %K (Figs. 5.2-5.3) was estimated by the bootstrap method, computing the 80% CL interval of 10,000 replicates, except for the two regions where none of the sampled individuals have haplogroup K ('Portugal coastal Early Neolithic' and 'Romania Late-Middle Neolithic'). Then the bootstrap method cannot be applied directly (because the error would be exactly zero, which is not reasonable), and thus we applied a different statistical method, explained in detail in Sec. 5.8.10. We have established the existence of the cline in 3 ways: linear regression (Fig. 5.2), interpolation map and Moran's I correlogram (Sec. 5.8.4).

5.4.3. **Analysis of K haplotypes**

We have applied several statistical and phylogenetic analysis to the K haplotypes found in the 9 Early Neolithic regional cultures: we have computed Tajima's D and Fu's F_S neutrality tests; analyzed the geographical variation in the haplotype diversity, mismatch distributions, and first principal component; correlated genetic and geographical distances through Mantel test; performed network analysis; and constructed a Bayesian Skyline Plot (Sec. 5.8.1). The obtained results show clear signs of a recent demographic and spatial expansion, in agreement with our assumption that haplogroup K spread with the Neolithic wave. These analyses have also shown as that, in principle, the regions displaying high values of %K are not the result of sampling individuals from a single family (see Sec. 5.8.1.2) Haplotype diversity).

5.4.4. **Space-time genetic simulations**

We use a rectangular grid of square cells that covers the European continent, the Near East and part of Asia and Africa, with each cell classified as inland, coast, mountain or sea [47]. We use cells of 50 km x 50 km, since 50 km is the value corresponding to the mobility per generation according to ethnographic data of preindustrial populations [182]. At each cell we can have individuals of three populations: farmers who *have* haplogroup K, P_N; farmers who do *not have* haplogroup K, P_X; and hunter-gatherers, P_{HG} (no hunter-gatherer has haplogroup K). Each population would in principle include several different haplotypes, but since we are not interested in the evolution of any individual haplotypes, for simplicity the model used in the main paper does not consider any lower level

subgroups. Below we describe the most important processes of the model, but we include a more detailed description in Sec. 5.8.5.

Initial conditions. We applied the initial condition that at 8,233 yr BCE, the date of Ras Shamra (the oldest PPNB site in Syria from previous simulations studies [47]), all of the grid was empty of farmers except the cell that contains this site. In this cell, we set at 8,233 yr BCE the hunter-gatherer population density to zero, and the farmer population density to its saturation value ($P_{F\,max} = 3200$ individuals/cell, from ethnographic data [47, 89]). The PPNB Syrian archaeological and genetic data have different times and locations (the archaeological data is dated at 8,233 yr BCE and the genetic data at 7,258 yr BCE). For this reason, we have to set the %K at the cell containing Ras Shamra by trial and error so that the simulation yields the adequate value of the %K at the time and location of the genetic data in Syria (see details in Sec. 5.8.7). In all grid cells (except for the initial one), the hunter-gatherer population is initially set at its saturation value ($P_{HG\,max} = 160$ individuals/cell, from ethnographic data [89]), assuming that none of them has haplogroup K (see the Sec. 5.1 and Sec. 5.8.2).

Defining a generation as the mean age of the parents at the time one of their offspring is born (not necessarily the first), in simulations we use the mean value $T = 32$ yr obtained from ethnographic data [93]. Let t stand for the number of generations elapsed since the beginning of the simulation (8,233 yr BCE). For $t = 1,2,3$ … we apply the following cycle of 3 steps (changing their order would yield the same results):

1) Dispersal. At each cell, we update the values of P_N and P_X by computing how many farmers of both kinds arrive at the cell from other cells. We do this, as in previous work [47, 94, 182], with a simple model in which, for each cell, a fraction p_e (which is called the persistence in demography) of the population of farmers (independently of their genes) stays at the cell, and a fraction $(1 - p_e)$ relocates to the four nearest neighbor cells, each receiving a fraction $(1 - p_e)/4$. We use the mean value $p_e = 0.38$ obtained from ethnographic data [182]. We expect that including a set of distances and probabilities would lead to similar results [3]. If one or more of the nearest neighbors are mountain cells, they cannot receive population and each of the remaining neighbors receives a higher fraction. If one or more neighbors are sea cells, the corresponding fraction of the population (that would move there) travels by sea, and is equally distributed among coast cells that can be reached by sea in straight lines of up to 150 km (this is the adequate distance to obtain agreement with archaeological data, as seen in Sec. 5.8.6, and in previous work [47]). We do not update the number of HGs in each cell due to their dispersal, because exchange of HGs between saturated cells has no effect (since all HGs lack haplogroup K) and we assume that they do not disperse appreciably into cells in which their number has been lowered due to cultural transmission (see step 2 below).

2) Cultural transmission. This is the only step that was not included in our previous non-genetic simulations on a real map of Europe [47], because they considered only purely demic models. There are 3 modes of cultural transmission [52]. Vertical transmission is due to interbreeding (i.e., cross-matings between farmers and HGs). Horizontal (oblique) transmission is due to learning of agriculture by HGs from farmers of the same (the previous) generation. The latter two modes can be combined in a single mathematical model, namely horizontal/oblique transmission [3]. Here we shall consider only vertical transmission for simplicity, but we would reach the same conclusions if we considered, instead, any combination of vertical and horizontal/oblique transmission (Sec. 5.8.9).

After dispersal, in each cell there is a population of P_{HG} hunter-gatherers and $P_N + P_X$ farmers. To determine the population numbers of the new generation, we have to compute the matings that take place between and within those 3 population groups, and then apply the reproduction step. We assume that children of cross matings between farmers and HGs are farmers, in agreement with

ethnographic observations [139]. The number of cross matings between HGs and each group of farmers is [94]

$$couples\ HN = \eta \frac{P_{HG} \cdot P_N}{P_{HG} + P_N + P_X}, \tag{5.1}$$

$$couples\ HX = \eta \frac{P_{HG} \cdot P_X}{P_{HG} + P_N + P_X}, \tag{5.2}$$

where $P_{HG} + P_N + P_X$ is the total population present at the cell, and parameter η is the intensity of interbreeding [94]. The case $\eta = 1$ corresponds to random mating. The case $\eta > 1$ corresponds to more cross matings than under random mating [94], which is not realistic for farmers and HGs according to ethnographic data [208, 209] (moreover, $\eta > 1$ can lead to negative population numbers [94]). Therefore, in practice $0 \leq \eta \leq 1$.

From Eqs. (5.1)-(5.2) it is very easy to find the number of individuals P'_{HG}, P'_N, and P'_X who do not take part in HN neither NX matings. We can use them to compute the number of matings between *farmer* individuals of different genetic groups (i.e., between populations P'_N and P'_X) by using again vertical cultural transmission theory and taking into account we have no reason to assume that farmers of a genetic group (i.e., with or without haplogroup K) will have a preference for (neither against) mating with farmers of the same genetic group. Thus we apply random mating ($\eta = 1$) [94] for matings between farmers,

$$couples\ NX = \frac{P'_N \cdot P'_X}{P'_N + P'_X}. \tag{5.3}$$

3) Reproduction. We apply the following rules. (i) Each couple will have $2 \cdot R_{0,i}$ children, because $R_{0,i}$ (the net fecundity) is computed per parent and there are two parents per mating ($i = F, HG$). Ethnographic data indicate that the children of cross matings with one HG parent are farmers [139, 208], thus we use $R_{0,HG}$ for HH matings and $R_{0,F}$ for HN, HX, NN, XX and NX matings. If the number of individuals computed for some population group, cell, and time step is larger than its corresponding maximum ($P_{F,MAX}$ or $P_{HG,MAX}$), then we set it to the corresponding maximum value. We expect that a logistic model would yield similar results. In our simulations we use $R_{0,F} = 2.45$ [183], indicating that after a generation, the size of the new population is 2.45 times the size of the parent population. We assume that $R_{0,HG} = 1$, i.e. that the HG populations have reached a stationary state and they do not grow in number (not even after some HGs mate into the farming community, because converted HGs will still need part of the cell space after they become farmers); we do not expect our conclusions to change for other reasonable values of $R_{0,HG}$. (ii) For each kind of mixed genetic mating (HN and NX), in our simplest model we assume that the mother belongs to P_N in 50% of the matings, whose children will also carry haplogroup K since mtDNA is inherited from the mother (i.e., a 50% of the total offspring of mixed genetic matings will belong to P_N). A more complicated model, assuming that mothers in HN and HX matings are always HGs (which is closer to ethnographic observations [208]) yields very similar results (Sec. 5.8.11).

All the steps in the model are computed using real values for the population numbers. If we used a stochastic procedure to approximate them to integer values (at every cell, iteration, and process step), we expect that in average we would obtain the same results. We run our simulation program for 200 iterations (generations of 32 yr) for each set of parameter values, so that it covers the time from the start of the spread (Syria, 8,233 cal yr BCE) until the latest genetic data in the database (Sweden, 2,825 cal yr BCE; Appendix A Data S3). At each iteration we compute the number of HG, N and X individuals

at each cell and record the latter two, so that we can compute the simulated %K (namely, $\frac{P_N}{P_N+P_X} \cdot 100$) and compare it to the observed one from the reported mtDNA data at each regional culture and its average date (Appendix A Data).

5.5. Acknowledgments

This work has been partially funded by Ministerio de Economía, Industria y Competitividad (Grant FIS-2016-80200-P), Fundación Banco Bilbao Vizcaya Argentaria (Grant NeoDigit-PIN2015E), and an Academia award from the Catalan Institution for Research and Advanced Studies (to J.F.).

5.6. Author contributions

JF conceived the research and devised the statistical method in Sec. 5.8.10. NI and VLR wrote the simulation codes. VLR compiled the genetic database and prepared figures. NI performed the genetic analyses in Secs. 5.8.1 and 5.8.4. JF, NI and VLR wrote the paper and the Supplementary Information Texts (Sec. 5.8) and Data (Appendix A).

5.7. Competing financial interests

The authors declare that they have no competing interests.

5.8. Supplementary Information Texts

5.8.1. Text S1. Analysis of K haplotypes. Signs of spatial expansion

Section 5.8.1 is devoted to independent analyses that confirm some claims made in our main paper. Therefore, the reader interested in detail on the data or the model used in the main paper can jump directly to Sec. 5.8.2 or S5, respectively.

As explained in the Introduction of the main paper, and as we shall see in Sec. 5.8.2 below, haplogroup K was virtually absent in pre-Neolithic Europe, whereas numerous Early Neolithic farmers carry haplotypes belonging to this haplogroup. This leads to the hypothesis that haplogroup K spread demically with the Neolithic wave, and we have applied this hypothesis to build the simulations reported in the main paper. Note that while the Neolithic spread could have been partially cultural (in the sense that hunter-gatherers could have contributed K individuals to farmer populations), the spread of haplogroup K, if absent in the local hunter-gatherer populations, must have been purely demic (in the sense that hunter-gatherers did not contribute K haplotypes to farmer populations). Therefore, if haplogroup K spread demically with the Neolithic front, one would expect to find signs of demographic and spatial expansion in the diversity of K haplotypes found in the Early Neolithic populations.

Our database includes 56 Early Neolithic individuals presenting mitochondrial haplotypes identified as belonging to haplogroup K (see Appendix A Data S1 and Data S2). For 55 of these individuals, at least part of the HVS-I region had been sequenced and the sequences were available in the respective sources cited in Appendix A Data S1 (the exception is sample deb29II, from the region '5 Western Germany LBK', for which the sequence for the HVS-I region could not be determined [78]). The range

shared by all sequences spans nucleotide positions 16106-16390. Because the HVS-II region is not sequenced for all individuals, and different authors test different coding region SNPs, in this section we shall apply our analyzes over this HVS-I range (see Appendix A Data S7).

Therefore, in this section we shall study only the 55 Early Neolithic individuals in Appendix A Data S1 identified as presenting haplogroup K and for which the HVS-I region has been sequenced and apply some statistical and phylogenetic analyses at the haplotype level to provide additional support to the hypothesis that haplogroup K spread demically with the Neolithic front. Our results will show clear signs of a recent expansion. Thus, given that haplogroup K was apparently absent from pre-Neolithic populations, and that there is no archeological record of other large demographic movements close in time to our data, the most reasonable conclusion from our results is the assumption made in our main paper that haplogroup K spread into Europe with the Neolithic front.

5.8.1.1. Tajima's D and Fu's F_s neutrality tests

We have analyzed the 55 sequences of Early Neolithic individuals with haplogroup K using Arlequin 3.5 [210], and computed the results for two neutrality tests: Tajima's *D* [211] and Fu's F_S [212]. For nucleotide positions 16106-16390 we can identify 12 different haplotypes (see Table 5.1 below), and we obtain significantly negative values for both statistics, $D = -2.10171$ and $F_S = -11.69788$. A negative value of *D* can be a result of selection, but it can also be due to a recent bottleneck or a process of population growth [211], and a negative value of F_s is often used as indicative of population expansion [212, 213]. Therefore, those results would be consistent with a recent process of demographic expansion [169, 211, 214], which is to be expected if we assume that haplogroup K spread demically with the Neolithic, so that farming populations underwent a process of demographic expansion.

Haplotype	HVS-I polymorphisms (16106-16390) [a]	Number of individuals	Regions found [b]
H01	T16224C T16311C	37	1, 2, 3, 4, 5, 6, 7
H02	T16311C	2	1
H03	T16224C T16311C C16366T	3	1
H04	T16224C T16311C G16290A	1	2
H05	T16189C T16224C T16311C	4	2, 3
H06	A16166G T16224C T16311C	1	3
H07	T16172C T16224C T16311C	1	3
H08	T16224C C16261T T16311C	1	3
H09	T16224C T16249C T16311C	2	4, 5
H10	T16209C T16224C T16311C	1	4
H11	T16224C T16311C G16319A	1	4
H12	T16224C T16311C T16362C	1	11

Table 5.1 K haplotypes in Early Neolithic regions. [a]Polymorphisms relative to rCRS [215]. [b]Region numbers correspond to the geographical region labels used in all figures and the Appendix A.

The mitochondrial region that we have used may in principle present a limitation as it does not include the polymorphic site at 16093, often used to discriminate K1a sub-haplogroups. Therefore we have repeated the analysis over the HVS-I range 16056-16390 for the 46 samples such that this range is sequenced (thus we have had to leave out of the analysis the 6 samples from '1 Syria PPNB', sample I0727 from '2 Anatolia', and samples 1CH0102 and CSA152223 from '6 North-Eastern Spain Cardial'). Using this reduced dataset we now obtain a value of Tajima's *D* not significantly different from zero at the 95% CL, which would indicate neutrality of mutations, while Fu's F_s is still significantly negative

($F_S = -7.90046$). Because Fu's statistic is especially sensitive to processes of population expansion [212, 213], and since Tajima's *D* is not positive in this analysis and neither in that in the previous paragraph, these results reinforce the proposal in our main paper that the observed diversity is the result of a demographic expansion, rather than of any possible process of background selection.

5.8.1.2. Haplotype diversity

The analysis of the evolution of haplotype diversity [216] can also help in identifying processes of population expansion. Here we shall analyze the evolution of haplotype diversity over space to identify signs of geographical expansion, that is, a decrease in the diversity with distance from the assumed source (see e.g. reference [217]).

Because of the low number of Early Neolithic individuals with haplogroup K in our database, to increase the significance of the samples in this section (and in the following sections) we have pooled the samples from geographically close regions (namely, regions 4-5 in Germany, and 6-7 in Iberia). We have computed the regional haplotype diversity indices using Arlequin 3.5 [210]. The results (Fig. 5.4) show a general decreasing trend with distance from Syria (with the exception of Anatolia, which shows relatively low haplotype diversity), and are thus indicative of a geographical spread from Syria [217]. This reinforces the conclusion in our main paper that haplogroup K spread.

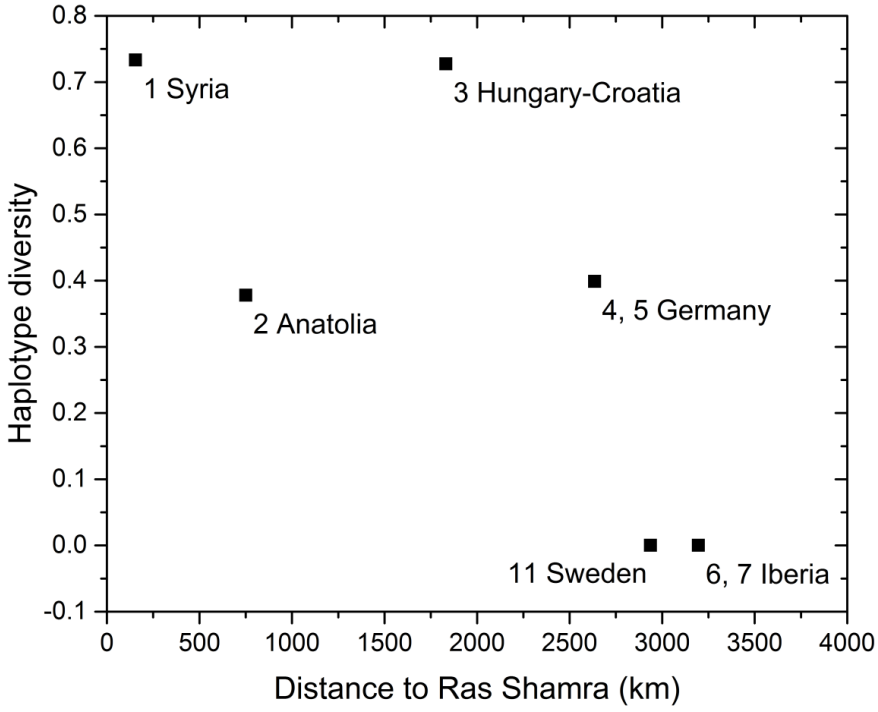


Figure 5.4 Haplotype diversity versus distance for Early Neolithic regions. This index shows a global decreasing trend, in agreement with a process of spatial expansion.

The low haplotype diversity found in Anatolia is in fact consonant with the fact that most samples in Anatolia present haplotype H01 (Table 5.1) and could in principle indicate that the samples correspond to a single family unit. However, upon examining the source, this does not seem to be the case for three reasons: (i) the samples correspond to two different sites; (ii) the analysis of the whole mtDNA sequences performed by Mathieson *et al*. [80] does not seem to indicate that the individuals are directly related; and, more conclusively, (iii) they display different subclades of haplogroup K.

Therefore, the low Anatolian diversity is probably due to the short nucleotide range that we are able to analyze in this study, as well as to sampling hazards. Similarly to Fig. 5.4, we expect the data from Anatolia will not follow the general trend in any of the regional analyses of mtDNA sequences performed in the next subsections. Below we shall find that this is indeed the case.

5.8.1.3. Mismatch distribution

The distribution of nucleotide site differences between pairs of individuals in a population can provide evidence of past demographic expansions undergone by this population [218]. Likewise, population range expansions can also leave similar traces in the distribution of pairwise genetic differences [89, 219, 220], with spatial signatures that can vary depending on the demic or demic-cultural nature of the expansion process [89].

Firstly, we have plotted the distribution of genetic differences including all 55 Early Neolithic individuals with haplogroup K, using the shared range of the HVS-I region 16106-16390. The result is shown in Fig. 5.5a. Because of the limitation of the analyzed range, the maximum number of differences if low, but the plot shows a distribution with a maximum close to zero differences, which would be consistent with a recent demographic or spatial expansion of the considered population (early Neolithic farmers with mtDNA haplogroup K) [218, 220].

As we have explained at the beginning of this section (Sec. 5.8.1), on the basis of the genetic evidence, in our simulations in the main paper we have assumed that haplogroup K spread demically with the Neolithic wave (because it was absent in European hunter-gatherer populations). In that case, one would in principle expect differentiated mismatch distributions of K haplotypes at different regions, with a maximum closer to zero in the case of populations located further away from the source, as shown by means of simulations by Currat and Excoffier (Fig. 4a-b in their results) [89]. In Fig. 5.5b-f we show the mismatch distributions for different geographical regions (to increase the significance of each sample, we have pooled geographically close regions 4-5 and 6-7 as in the previous subsection; Sweden cannot be analyzed here as it has only one individual with haplogroup K). The results, while clearly limited by the low number of individuals and analyzed nucleotide positions, do show a trend in which the maximum identified in Syria (Fig. 5.5b) moves closer to zero at geographically distant locations (again with the exception of Anatolia, which shows a peak closer to 0 differences than expected for a region close to the source). We conclude that the general trend observed in the distributions is as expected from previous simulations [89] and can thus be interpreted as a result of a recent geographic expansion of individuals with haplogroup K. This agrees with our assumption that haplogroup K spread demically with the Neolithic front.

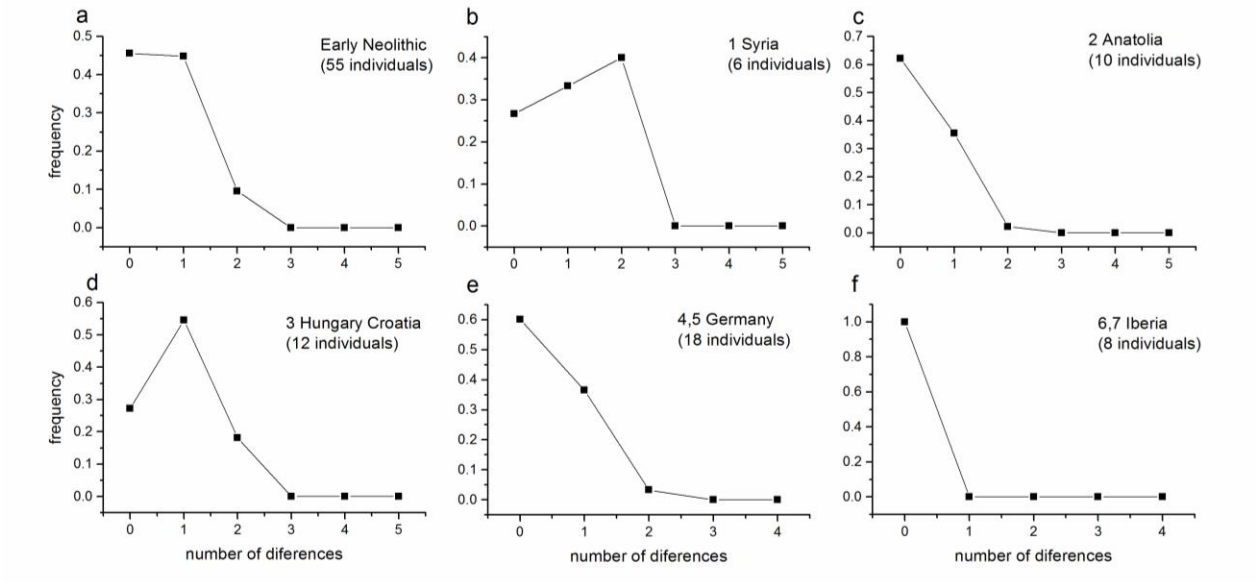


Figure 5.5 Mismatch distributions for K haplotypes identified in all Early Neolithic samples (upper left) and in specific regions. The distributions have been obtained from mtDNA sequences for the HVS-I region at nucleotide positions 16106-16390. (a) includes all 55 early Neolithic individuals with haplogroup K, whereas (b)-(f) correspond to regional samples.

5.8.1.4. Mantel test

It is well-known that a process of geographic expansion leads to a strong increase of genetic distance with increasing geographic distance [221, 222]. For this reason, we have computed the pairwise genetic distance F_{ST} between the Early Neolithic regional cultures (considering only K haplotypes) and performed a Mantel test [223, 224] to evaluate the correlation between genetic and geographic distance matrices [221, 225, 226]. Genetic distances and Mantel tests were computed with Arlequin 3.5 [210] performing 10,000 permutations. In order to increase the significance of each sample we have pooled the 55 Early Neolithic individuals presenting haplogroup K into six geographic areas (as done in the previous subsection): Syria, Anatolia, Hungary-Croatia, Germany, Iberia, Sweden.

Surprisingly, the results of applying a Mantel test to the genetic and geographic distance shows a very low matrix correlation value $R = 0.15$. Examining the data, the reason for this low value can be partially attributed to the fact that there is only a single K haplotype in Sweden, which in turn differs from all other K haplotypes analyzed, thus leading to a very high value of the genetic distance to other close regions. The sample from Sweden is also dated considerably later than the other samples (see Fig. 5.1 in the main paper), so the genetic distance could be due not only to geographical distance, but also to temporal distance (indeed, applying a Mantel test to genetic and temporal distances yields a much better correlation value $R = 0.67$). For this reason we have computed anew a Mantel test for genetic and geographic distances leaving Sweden out of the analysis, which leads an increases value of the correlation between matrices, $R = 0.45$.

In Fig. 5.6 we have plotted the genetic versus geographic distances to Syria, in order to visualize the correlation between both distances [225]. This plot shows that Anatolia diverges considerably from the overall behavior, similarly to the observation from the previous subsection where the mismatch

distribution for Anatolia also diverged from our expectations. Thus we applied a Mantel test without Anatolia (nor Sweden), which leads to a much higher correlation value $R = 0.88$.

Therefore, we see that there is a spatial correlation with genetic distances, although the results when considering all regions are affected by the very late date for the sample in Sweden, and by the K samples from Anatolia, which seem to present a higher divergence than would have been expected. As mentioned when analyzing the haplotype diversity, this exception may be due to the low number of individuals and analyzed nucleotide positions in the data available at present.

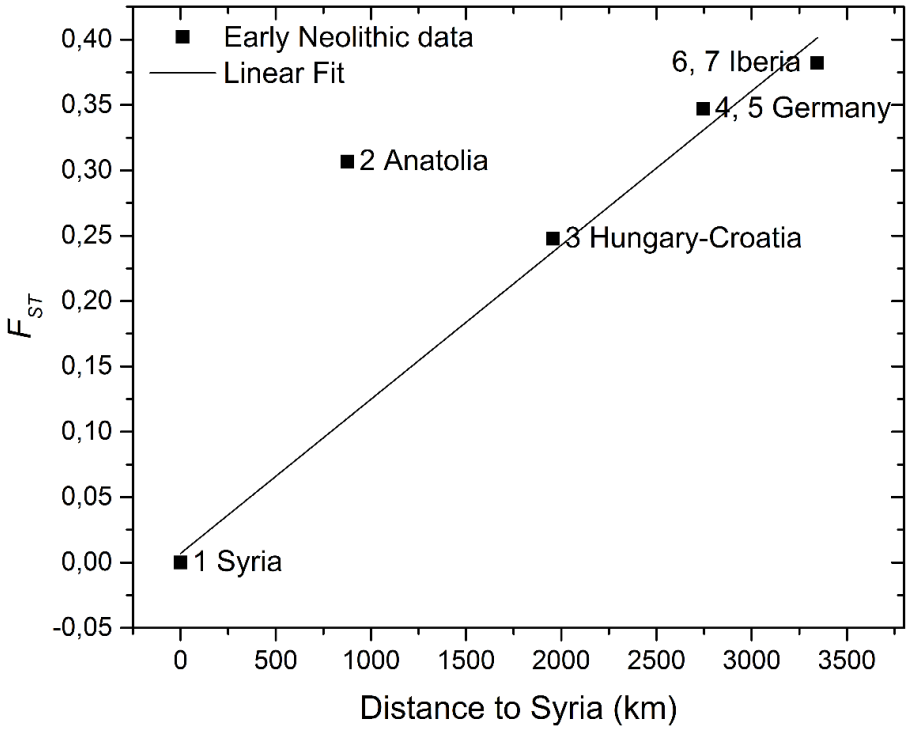


Figure 5.6 Genetic distances to the Syrian population versus geographic distances for Early Neolithic regions. The line corresponds to the linear fit without region 2 Anatolia. Note that these data correspond to the first column of the matrices of genetic (F_{ST}) and geographic distances used in the Mantel tests (Sweden is not included).

5.8.1.5. PCA Analysis

Alongside the Mantel test, we can also test the correlation between genetic and geographic distances by performing Principal Component Analysis (PCA) on the K haplotypic data. We have performed PCA between groups using PAST 3.15 software [227] for the different geographical regions (as in the previous subsections, we have pooled the data from geographically close regions).

We find that the first principal component (PC) explains a 63% of the variability between groups (the second PC explains a 22% of the variability), so below we plot the first PC against distance (Fig. 5.7). We see that, similarly to the results obtained above, there is a very clear spatial correlation between Syria, Hungary-Croatia, Germany and Iberia, while Anatolia (region 2) and Sweden (region 11) fall clearly out of this trend. There is a clear overall correlation between genetic differentiation and

distance (Fig. 5.7), and this is consistent with the involvement of haplogroup K in the Neolithic demic flow.

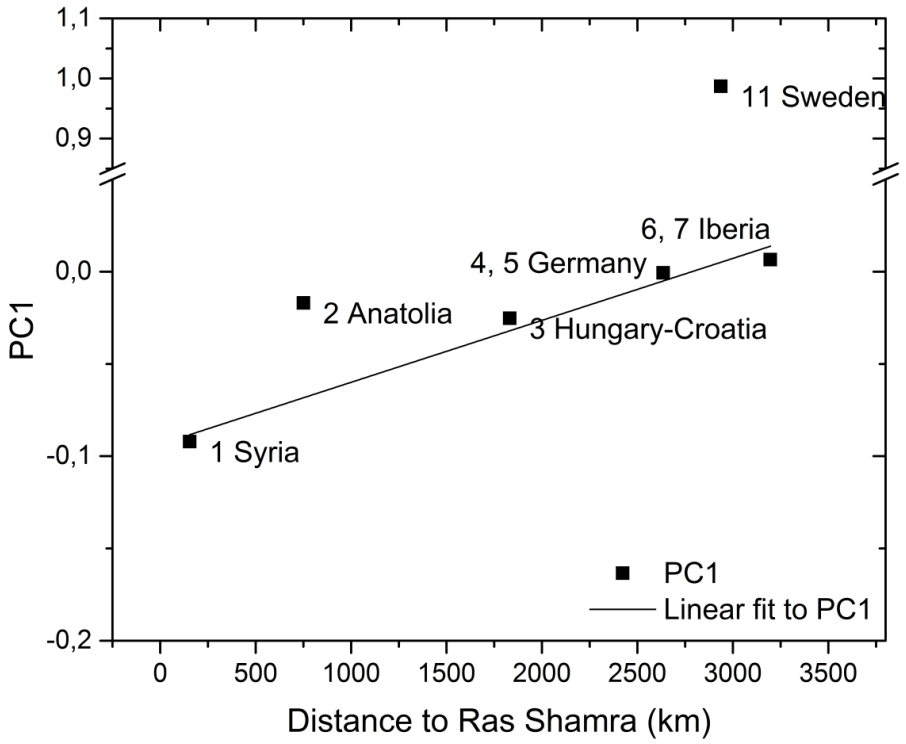


Figure 5.7 Variation of the first principal component (PC1) with distance from Ras Shamra. The line corresponds to the linear fit obtained excluding regions 2 (Anatolia) and 11 (Sweden).

5.8.1.6. Network analysis

When a population undergoes an expansion process, it has been shown that phylogenetic network analysis leads to star-shaped genealogies [228]. Figure 5.8 shows the median-joining network obtained with Network 5 software [229] (www.fluxus-engineering.com) for the 55 Early Neolithic HVS-I sequences (nucleotide positions 16106-16390). The obtained results are clearly star-shaped, although to reinforce this observation we have computed the star index introduced by Torroni *et al*. to evaluate the *starness* of a phylogeny [65]. This index is defined as the relative frequency of pairs of sequences that coalesce at the assumed root (in our case, haplotype H01), and a value >0.95 is considered to reflect a highly star-like group [65, 230]. From the data in Table 5.1 we obtain that only 11 of the 1485 possible pairs of sequences do not coalesce at the root, thus the star index for the early Neolithic haplogroup K is 0.99. This indicates that we have indeed a very star-like phylogenetic network (in agreement with a process of population expansion), and that haplogroup K was involved in the Neolithic demic flow (as assumed in our main paper).

Figure 5.8 also provides a supplementary visualization to Table 5.1 above, which shows clearly that the most abundant haplotype is H01. This haplotype H01 is present in all regions but Sweden, while all of the other haplotypes are present in only one or two close regions (see Table 5.1). Therefore, haplotype H01 would have been carried on along the whole expansion, while other haplotypes might have appeared locally but not spread in the process of spatial expansion.

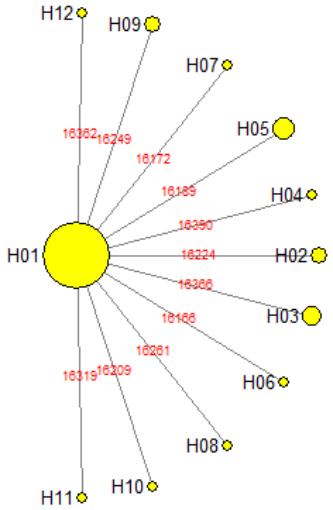


Figure 5.8 Median-joining network of K haplotypes present in Early Neolithic cultural regions. The nodes correspond to the haplotypes listed in Table 5.1 and their sizes are proportional to the number of individuals. The mutated nucleotide positions are indicated at the links.

5.8.1.7. Bayesian Skyline Plot

In this section, we have applied Bayesian coalescent inference to study the variation in time of the effective population of individuals bearing K haplotypes at the Early Neolithic front. We have generated a Bayesian skyline plot (BSP) [231] for the Early Neolithic HVS-I sequences (positions 16106-16390) corresponding to individuals carrying K haplotypes, each one dated with its calibrated date (see Appendix A Data S7). The BSP was generated using BEAST 2 [232] and Tracer 1.6 [233]. The Markov chain Monte Carlo (MCMC) samples were based on a run of 40,000,000 generations, sampled every 40,000 generations, and with the first 10% discarded as burn-in. We used a JC69 substitution model (although using a HKY yields very similar results) and a strict clock with a mutation rate 1.62×10^{-7}, as reported by Soares *et al*. [234] for the HVS-I region.

Figure 5.9 shows the BSP obtained for the Early Neolithic individuals presenting haplogroup K. Because the individual with haplogroup K in Sweden is dated about 2,000 yr later than the other Early Neolithic data (see Appendix A Data S7), we have not included Sweden in the results shown in Fig. 5.9. Figure 5.9 shows that the effective population size remains mostly stationary with a decreasing trend throughout the considered period. Whereas the Neolithic spread is associated with a process of population growth, we have seen in the main text (and we shall further discuss in Sec. 5.8.4) that the percentage of the population carrying K haplotypes decreased at the Neolithic front, thus a stationary evolution of the population size of haplogroup K is a reasonable result. In addition, while rapid population growth processes are often related to the retention of genetic diversity [235, 236, 237], stationary populations (Fig. 5.9) have been related with a loss of haplotype diversity [238], in agreement with our observations from Fig. 5.4.

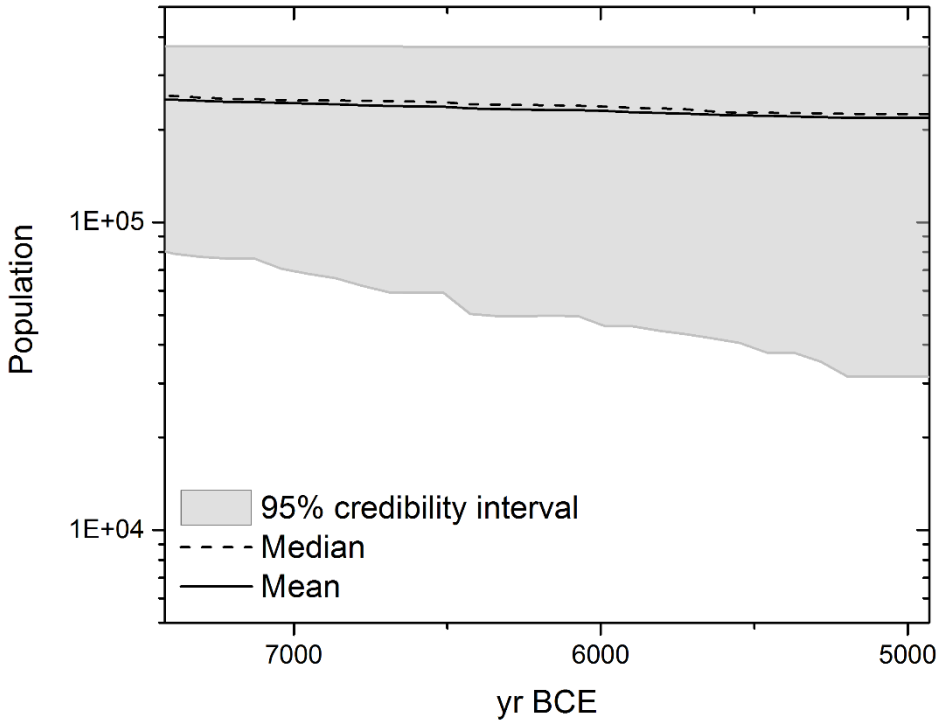


Figure 5.9 Bayesian skyline plot showing the evolution of the effective population size of K haplotypes in Early Neolithic groups in through time. Sweden is not included because its single date is from 2,000 yr after the youngest extreme of the range in this figure. The solid and dashed lines indicate, respectively, the mean and median population sizes, and the shaded region corresponds to the 95% credibility interval.

5.8.2. **Text S2. Mesolithic samples with haplogroup K**

As explained in the main paper: (i) haplogroup K has been found in ancient farmers in many sites of Europe, as well as in Anatolia and the Near East; (ii) in contrast, no Western neither Central European hunter-gatherer has been found so far with haplogroup K before the Neolithic period; (iii) there are very few cases of hunter-gatherers with haplogroup K. For reasons (ii) and (iii), it is very reasonable to consider haplogroup K as virtually absent in pre-Neolithic Europe. Still more, there are even reasons to disregard the very few cases of hunter-gatherers with haplogroup K mentioned in point (iii). We explain these reasons in this section. Up to date, a total of 8 Mesolithic individuals with haplogroup K have been found. One is from Germany, four from Sweden, two from Greece and one from Georgia. We discuss them in turn.

One hunter-gatherer (*OstorfSK28a*) with haplogroup K (no subclade was reported by Bramanti *et al*. [203]) was found in Ostorf, a Mesolithic site in northern Germany, and dated 3,200 cal BCE. However, as noted by Bramanti *et al.* [203], it is very remarkable that Ostorf is a Mesolithic enclave surrounded by farmers (of the Funnel-beaker culture). Moreover, Ostorf is precisely the single hunter-gatherer site where individuals with non-U mtDNA haplogroups were found [203]. Thus, it is reasonable to consider the possibility that haplogroup K was introduced in Ostorf by interbreeding with farmers.

Four hunter-gatherers (*Ire9, Fri28, GE76, Vis7B*) have been found in Sweden (Pitted Ware culture, PWC) with subclades K1a and K1a1, and dated 3,200–2,400 cal BCE [168, 176]. Despite its hunter-gatherer economy, the PWC overlapped chronologically with farmers during almost a millennium, first

of the Funnel-beaker culture (*Trichterbecherkultur*, TRB) and later of the Battle Axe complex, a variant of the Corded Ware culture [239, 240]. This is why some authors refer to the PWC as 'Neolithic' hunter-gatherers [177]. Thus it is again reasonable to consider the possibility that this small sample of hunter-gatherers with haplogroup K (4 of 32 PWC individuals) is due to interbreeding with contemporaneous farmers living in the same region.

Two hunter-gatherers (*Theo1* and *Theo5*) displaying subclade K1c were discovered in Theopetra, a site in Thessaly (Greece) and dated 7,605–6,771 years BCE [171]. However, subclade K1c (as well as subclades K2b and K2c) has been never found among Neolithic farmers to date. Thus these two Mesolithic individuals do not affect the subclades of haplogroup K that were presumably introduced into Europe by the Neolithic population wave of advance.

Similarly, a hunter-gatherer (*Satsurblia*) from Georgia (associated with the Epigravettian culture) has been dated 11,380–11,130 cal BCE [204] and displays the K3 subgroup, which has been never found among Neolithic farmers to date.

In view of these considerations, current evidence makes it very reasonable to believe that haplogroup K or, more precisely, the subclades of haplogroup K that have been found in European Neolithic individuals (see Appendix A Data S1 for the complete list), were absent in Europe before the spread of farming, and were introduced there by incoming farmer populations of Near Eastern origin.

5.8.3. **Text S3. Neolithic individuals not included in the study**

In this work, we have gathered a database of all individuals from farming cultures dated between 8,000 and 3,000 calibrated years BCE for which the mtDNA haplogroup have been reported in the literature. We have grouped these individuals into regional cultures according to their geographical and cultural closeness, but we have only selected for further analysis (Fig. 5.1) the 26 regional cultures with more than 2 individuals (Appendix A Data S1). Therefore, we have discarded 5 individuals from the database. In particular, we discarded 'Spain (Valencia and Alacant)' with only two individuals [173], and all data from Greece, because the one Early Neolithic individual is dated about 2,000 yr earlier than the two Late Neolithic individuals, and therefore they cannot be considered a single group (Appendix A Data S1).

Very recently, the first mtDNA data from ancient farmers in the southern Levant (Jordan and Israel) have been reported [81]. As mentioned in the main paper, we have not included them. The reason is that haplogroup K has been found in only 23% (3 of 13) PPNB/C individuals [81], and this is substantially lower than the value 40% that we obtain for the Syrian PPNB sites [165]. If future studies (based on larger databases) confirm a low %K in the southern Levant, it may have several causes. One possibility is simply that, as suggested by the genetic analyses by Lazaridis *et al.* [81], the ancient farming population from the southern Levant did not lead to the Early Neolithic populations in the Near East and Europe. A second possibility is that a drift effect could have increased the %K during the spread of the Neolithic from the Southern Levant to northern Syria. This second possibility is an open issue and would, in any case, require a substantially more complicated model (based on additional assumptions), which is out of the scope of the present paper. Thus we consider ancient mtDNA data from Syria, Anatolia and Europe, which (as we have seen) do show a fairly gradual spatial decrease (i.e., a cline) in the %K, in agreement with our simple model. Admittedly, we expect that future work will lead to more general models that can describe more complicated clines.

5.8.4. **Text S4. Geographic cline of haplogroup K**

Similarly to Sec. 5.8.1, this section presents some analyses that are independent of the method used in the main paper but reinforce an important claim made in our study.

In the main text (Figs. 5.2-5.3) we have visualized the geographic cline of haplogroup K by representing its measured relative presence in different regions as a function of the great-circle distance to Ras Shamra (Syria), the oldest PPNB archaeological dating in reference [47] (the great-circle distance is the shortest distance between two points on the surface of a sphere; in this case, on the surface of the Earth). This representation is the most effective option to take into account the effect of low samples (since we take into account the whole 80% CL range, plotted as error bars) and to compare the simulation results with the measured results (e.g., Fig. 5.3). However, it might not be the most intuitive way to understand the distribution of haplogroup K throughout the European continent. In Fig. 5.10 we have represented the locations of the 9 Early Neolithic regions used in Fig. 5.3, labeled with the percentage of population presenting haplogroup K, which we have interpolated using Ordinary Kriging with the software ESRI ArcGIS 10.4. The interpolation results show clearly that there is a spatial gradient on the presence of K haplogroup, both along the Mediterranean as well as along the interior spread route.

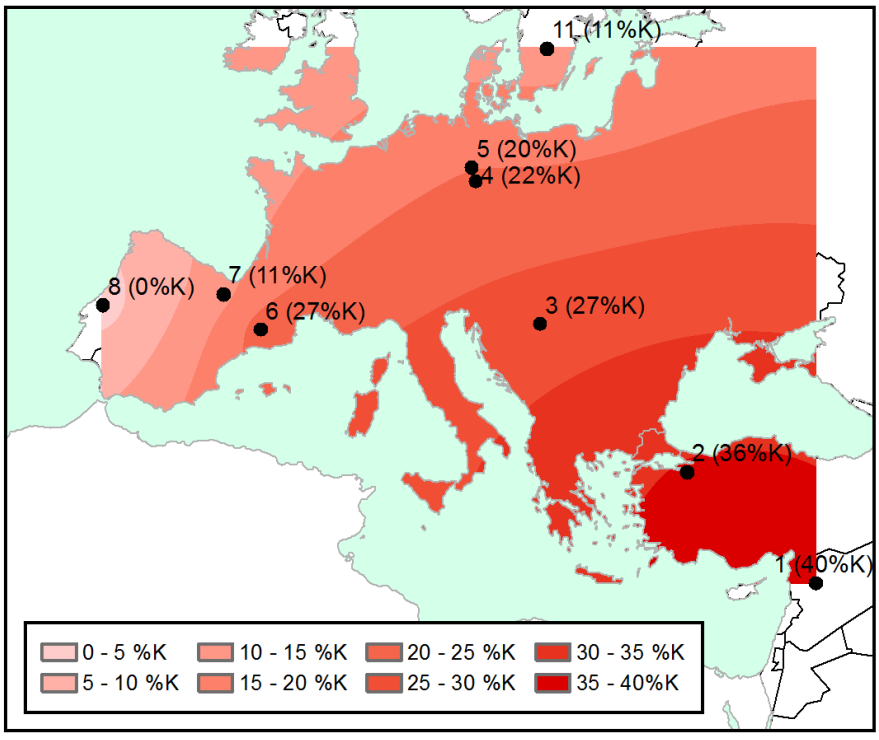


Figure 5.10 Spatial gradient of haplogroup K in Early Neolithic populations. Circles represent the location of the 9 Early Neolithic cultural regions shown in Fig. 5.3, labeled with the %K in each of them. The kriging interpolation shows the spatial decrease in the presence of K haplogroups away from Syria. Map created with ArcMap 10 and the Spatial Analyst 10 extension (http://desktop.arcgis.com/es/desktop/).

An alternative technique to detect the presence of a spatial cline is by studying the spatial autocorrelation of the data through a Moran's I correlogram [241], as has been previously done to

analyze geographic patterns from genetic data [242]. When the data display a spatial cline, the correlogram should show a decreasing behavior, with positive autocorrelation at short distances and negative autocorrelation at long distances [241, 242] (i.e., nearby points are similar whereas distant points differ). On the other hand, a random spatial distribution of observed values (i.e., a non-clinal pattern) would display a flat correlogram with an expected value of Moran's I given by (see reference [241], Eq. (13.6))

$$E(I) = -1/(N-1), \tag{S1}$$

where N is the number of data points (9 regions in our case, so $E(I) = -0.125$). Note that, for a random (thus non-clinal) spatial distribution of observed values, $E(I) \to 0$ if $N \to \infty$ [241].

Figure 5.11 shows the correlogram obtained with PASSaGE 2 [243] for the %K present at the same 9 Early Neolithic regions as in Fig. 5.3 and Fig. 5.10. We have grouped the great-circle distances between pairs of regions into 6 distance classes (in agreement with Struge's rule; equation 13.3 in reference [241]), chosen so that there is an equal (or nearly equal) number of observations per class. The correlogram is significant over the entire range of classes (P<0.005 Bonferroni corrected [241, 244]) and shows a clinal trend, as expected if there is a spatial gradient of the presence of haplogroup K [242]. Repeating the same computation but using 6 distance classes of equal width, also yields a significant cline (results not shown; P<0.05 Bonferroni corrected).

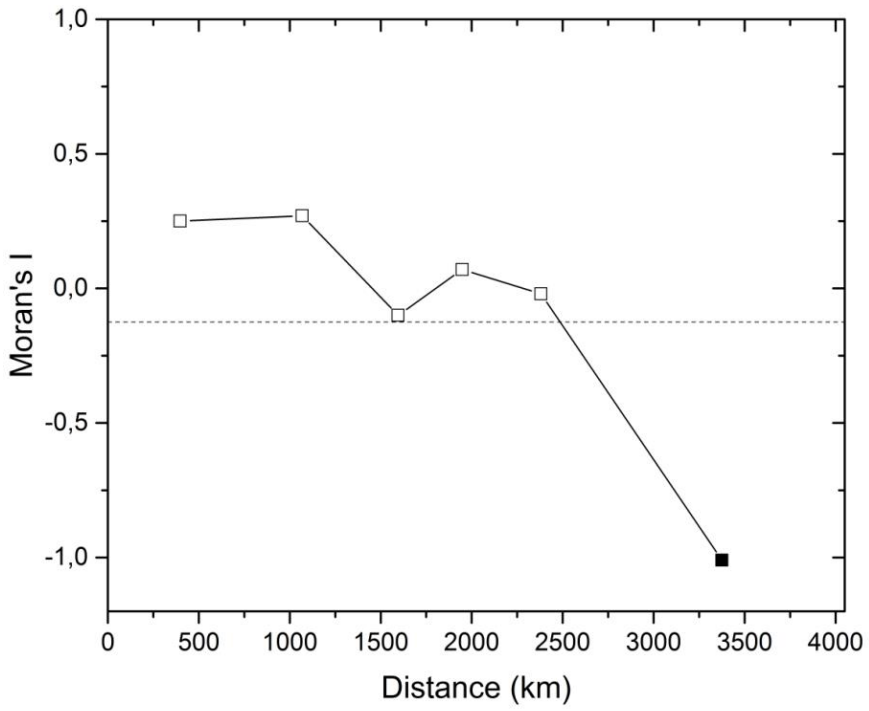


Figure 5.11 Spatial correlogram for the presence of haplogroup K in Early Neolithic cultural regions. The dashed line shows the expected value of I under a random (i.e., non-clinal) spatial distribution, $E(I) = -0.125$, from equation (S1) (see [241], Eq. (13.6)). Black dots correspond to class-specific significant values. The correlogram is significant over the entire range of classes (P<0.005 Bonferroni [241, 244]) and displays a clinal behavior.

Therefore, the results obtained here reinforce our conclusion from Figs. 2-3 that there is a spatial cline in the percentage of Early Neolithic farmers carrying haplotypes from haplogroup K.

5.8.5. Text S5. Mathematical details of the computational model

The Fortran code for the model used in the main paper, and described below, is available as Program S1 at the journal web or at http://copernic.udg.es/QuimFort/2017_08_07r__Program_S1.zip.

As explained in Sec. 5.4, the model runs on a grid of 50x50 km^2 square cells (180x120=18,360 cells). Elevation data from the SRTM30 near-global elevation model were used to determine the main type of terrain (inland, mountain, coast or sea) of each cell [47]. For coast cells, one of the four nearest neighbors must be a sea cell, while inland cells cannot have a sea cell as one of its nearest neighbors. Neolithic and Mesolithic individuals can only inhabit inland or coast cells. Each of these cells can have a maximum famer population of $P_{F\,max} = 3{,}200$ individuals/cell [47], which includes farmers with and without haplogroup K (this value was computed from the ethnographic data on the maximum density [89], 1.28 individuals/km^2, and the area of the cell, 2,500 km^2), and a maximum hunter-gatherer population $P_{HG\,max} = 160$ individuals/cell (obtained from the ethnographic maximum density [89], 0.064 individuals/km^2). Here we consider areas higher than 1,750 m above sea level as mountain barriers. However, the results are very similar changing the value of 1,750 m by other values, and also if neglecting mountain effects altogether, as previously observed for non-genetic simulations [47].

Each cell is assigned an initial population of farmers with haplogroup K, $P_N(x, y, t = 0)$, farmers who do not have haplogroup K, $P_X(x, y, t = 0)$ and hunter-gatherers $P_{HG}(x, y, t = 0)$, as follows. Initially, $P_{HG} = 0$, $P_N + P_X = P_{F\,max}$ at the cell with coordinates (112, 31) that contains Ras Shamra, the oldest PPNB site in Syria (the values of P_N and P_X will depend on the parameters used; see details in Sec. 5.8.7), and $P_{HG} = P_{HG\,max}$, $P_N = 0$ and $P_X = 0$ at all other cells. Given these initial conditions, the model updates each of the three populations (N, X, HG) at every iteration (generation of 32 yr [182]) $t = 1,2,3$ …, according to three steps: dispersal, interaction, reproduction (changing the order of these 3 steps would yield the same results). Note that at any instant, the total farming population per cell is given by $P_F(x, y, t) = P_N(x, y, t) + P_X(x, y, t)$. Each of the three populations considered in the model would comprise several different haplotypes, but since we are only interested in their results at the haplogroup level, we do not further subdivide the population. All computations are performed using real values, though we expect that, in average, we would obtain the same results if we used a stochastic procedure to approximate them to integers at each of the following three steps of the process.

5.8.5.1. Dispersal

Under the reasonable assumption that farmers have the same dispersal behavior independently of their mtDNA haplogroup, in this step we apply the following rules to each of both subpopulations.

Persistency. A fraction p_e of the subpopulation initially present at each cell remains in it (p_e is called the persistence in demography). The rest (fraction $1 - p_e$) moves to other cells, as follows. In the model we use the mean value $p_e = 0.38$ obtained from ethnographic data [139].

Land travel. The farmers that move from a cell (which may be inland or coast) can travel by land to some of its four nearest neighbor cells. We could consider a set of more than two inland travel distances (0km and 50km in our model) and their corresponding probabilities, with all distances and probabilities estimated from ethnographic data [3, 183], but this would require substantially more computer time, and we expect it would lead to similar results (so we consider only the characteristic

distance moved per generation according to ethnographic data, namely 50 km [139]). As said above, Neolithic populations can only settle on inland or coast cells (mountain cells cannot be inhabited and act as barriers that cannot be penetrated; sea cells cannot be inhabited either but allow individuals to travel by sea to other locations). Therefore, if none of the four nearest neighbors to an inland cell are mountains, each of the 4 inland or coast neighbors receives 1/4 of the population that relocates, i.e. a fraction $(1-p_e)/4$ of the population at the initial inland cell. If one of the neighbors is a mountain, it acts as a barrier, and no population will move to this cell; as a result, each inland or coast neighbor receives a fraction $(1-p_e)/3$. Similarly, if two of the nearest neighbors are mountains, each remaining inland or coast cell receives a fraction $(1-p_e)/2$. In general, the fraction of the population that moves to *each* inland or coast neighbor is given by

$$\frac{(1-p_e)}{(4-\#mountain\ neighbors)}. \tag{S2}$$

Sea travel. Consider population leaving a coast cell. If only one of its neighbors is a sea cell, the fraction of the population that would travel by land to this cell (according to equation (S2)) travels by sea to other coast cells. If the initial coast cell has two sea neighbors, the fraction of the population that travels by sea is twice the value given by equation (S2), i.e. the number of individuals that would travel by land to both sea cells. In general, the *total* fraction of the population that travels by sea from a given cell is

$$\frac{(1-p_e)\cdot \#sea\ neighbors}{(4-\#mountain\ neighbors)}. \tag{S3}$$

For example, if a coast cell has one sea neighbor, two coast neighbors and one mountain neighbor, according to equation (S2) each coast neighbor would receive a fraction $(1-p_e)/3$ of the population in the origin cell, and, according to equation (S3), an equal fraction $(1-p_e)/3$ would travel by sea. As another example, if a coast cell has two coast and two sea neighbors, according to equation (S2) each coast neighbor would receive a fraction $(1-p_e)/4$ of the population in the origin cell, while now a fraction $(1-p_e)/2$ would travel by sea, according to equation (S3).

Sea travel takes place in straight lines across the sea to other coastal cells within a given range. We select as sea-travel destinations all coastal cells within the sea-travel range (measured along straight lines), that can be reached following a linear route that crosses only sea cells; i.e. those coastal cells within line of sight across the sea (and within the maximum sea travel distance). Each possible destination receives an equal fraction of the population that travels by sea. Therefore, if there are for example 5 possible destinations, each one receives 1/5 of the fraction of the population that travels by sea, which is given be equation (S3). In the simulation we use a sea travel range of 150 km. See Sec. 5.8.6 for details on how we determined this range.

We do not update the number of HGs at each node due to their dispersal, because the exchange of HGs between saturated cells has no effect (since the HG population lacks haplogroup K) and we assume that they do not disperse appreciably into cells in which their number is lower than the saturation value (due to cultural transmission, see below).

5.8.5.2. Cultural transmission

After dispersal, in each cell there is a population of $P_{HG}(x,y,t)$ hunter-gatherers and a population of $P_F(x,y,t)$ farmers. As mentioned above, $P_F(x,y,t) = P_N(x,y,t) + P_X(x,y,t)$, with $P_N(x,y,t)$ the number of farmers who have haplogroup K and $P_X(x,y,t)$ the number of farmers who do not have haplogroup K. As mentioned in our main paper, for simplicity we consider only interbreeding (vertical transmission), but we would reach the same conclusions if we considered, instead, acculturation (horizontal/oblique transmission), or both interbreeding and acculturation (see Sec. 5.8.9 for a detailed justification of this point). Under vertical transmission, to determine the population that will conform the new generation, we have to compute the matings that take place between and within those 3 population groups, and then apply the reproduction step.

Cross-matings between cultural groups. We assume that children of cross matings between farmers and HGs are farmers, in agreement with ethnographic observations [52, 208]. The number of cross matings between HGs and farmers is then given by [94]

$$couples\ HF = \eta \frac{P_{HG}(x,y,t) \cdot P_F(x,y,t)}{P_{HG}(x,y,t) + P_F(x,y,t)}, \tag{S4}$$

where $P_{HG} + P_F = P_{HG} + P_N + P_X$ is the total population present at the cell, and parameter η is the intensity of interbreeding [94]. The value of the interbreeding parameter lies in the range $0 \leq \eta \leq 1$, with the case $\eta = 1$ corresponding to random mating ($\eta > 1$ would correspond to more cross matings that under random mating, which is not realistic for farmers and HGs according to ethnographic data [208, 209] and, moreover, $\eta > 1$ can lead to $P_{HG} < 0$ for $P_{HG} \ll P_F$ [94]).

Here we are interested in the genetics of the offspring. In order to compute this, we need to consider separately the matings of HGs and farmers who have (P_N) or not (P_X) haplogroup K. Therefore, we separate the number of matings given by equation (S4) into two terms,

$$couples\ HN = \eta \frac{P_{HG}(x,y,t) \cdot P_N(x,y,t)}{P_{HG}(x,y,t) + P_F(x,y,t)}, \tag{S5}$$

$$couples\ HX = \eta \frac{P_{HG}(x,y,t) \cdot P_X(x,y,t)}{P_{HG}(x,y,t) + P_F(x,y,t)}. \tag{S6}$$

Note that $couples\ HF = couples\ HN + couples\ HX$, since $P_F = P_N + P_X$.

Within each population, the number of individuals who do not take part in HN neither HX matings is given by

$$P'_{HG}(x,y,t) = P_{HG}(x,y,t) - couples\ HN - couplesHX, \tag{S7}$$

$$P'_N(x,y,t) = P_N(x,y,t) - couples\ HN, \tag{S8}$$

$$P'_X(x,y,t) = P_X(x,y,t) - couples\ HX. \tag{S9}$$

Cross-matings between genetic groups of farmers. Let us next compute the number of matings between *farmer* individuals of different genetic groups, i.e. between populations P'_N and P'_X. Again, we can compute the number of mixed genetic couples using vertical cultural transmission theory. However, we have no reason to assume that farmers of a genetic group will have a preference for

(neither against) mating with farmers of the same genetic group. Thus we apply random mating ($\eta = 1$) [94] for matings between farmers. Therefore, the number of NX matings is

$$couples\ NX = \frac{P'_N(x,y,t) \cdot P'_X(x,y,t)}{P'_N(x,y,t) + P'_X(x,y,t)}. \quad \text{(S10)}$$

Note that we are indeed dealing with an equation equivalent to equation (S4), although now the total population we are considering is just the farmer population that does not mate with HGs, i.e. $P'_N + P'_X$. This completes the computation of the numbers of all possible cross-matings.

***Matings within groups*.** All remaining individuals, i.e. those that do not mate with individuals of a different group, will mate with individuals of their group (individuals that do not mate are not explicitly considered, since their effect is already taken into account by the net reproduction rate used in the next step). In these cases, obviously there are 2 individuals of the same group per mating, and the corresponding numbers of matings are

$$couples\ HH = P'_{HG}(x,y,t)_{HG}/2, \quad \text{(S11)}$$

$$couples\ NN = [P'_N(x,y,t) - couples\ NX]/2, \quad \text{(S12)}$$

$$couples\ XX = [P'_X(x,y,t) - couples\ NX]/2. \quad \text{(S13)}$$

where we have taken into account that matings NX have 1 N individual and 1 X individual, so the number of individuals N (or X) in matings NX is equal to the number of couples NX.

5.8.5.3. Reproduction

Finally, we apply reproduction to compute the new populations at each node a generation later. To do so we set the following rules. (i) Each couple will have $2R_{0,i}$ children, because $R_{0,i}$ is computed per individual and there are two individuals per mating. However, the net growth rate $R_{0,i}$ is different for farmers than for HGs ($i = F, HG$). Ethnographic data indicate that the children of cross-matings with one HG parent are farmers [52, 208], thus we use $R_{0,HG}$ for matings in which both parents are HGs, and $R_{0,F}$ for HN, HX, NN, XX and NX matings. (ii) For each kind of mixed genetic matings (HN and NX), in our simplest model we assume that the mother is N in 50% of matings, i.e. that 50% of the children from genetic mixed matings have haplogroup K (because mtDNA is inherited from the mother, and thus only the offspring from mothers bearing haplogroup K will have this haplogroup). Classical cultural transmission theory [93] assumes that $R_{0,F} = R_{0,HG} = 1$ (no population growth) but this is not our case, because we are dealing with a population expansion of farmers, so their number increases and we used instead $R_{0,F} = 2.45$, obtained from ethnographic data [183]. Under assumptions (i) and (ii), the number of individuals of each population group the next generation is related to the numbers of matings as

$$P_{HG}(x,y,t+1) = R_{0,HG}[2 \cdot couples\ HH], \quad \text{(S14)}$$

$$P_N(x,y,t+1) = R_{0,F}[2 \cdot couples\ NN + couples\ NX + couples\ HN], \quad \text{(S15)}$$

$$P_X(x,y,t+1) = R_{0,F}[2 \cdot couples\ XX + 2 \cdot couples\ HX + couples\ NX + couples\ HN]. \quad \text{(S16)}$$

where the factor 2 before the number of couples HH, NN and XX comes from the fact that each of those matings leads, the next generation, to $2R_{0,i}$ individuals of the same group as their parents.

Similarly, the factor 2 in front of the number of couples HX takes into account that each such mating leads to $2R_{0,F}$ farmers of genetic type X (with haplogroups different than K) the next generation. In contrast, each of NX or HN matings leads to $R_{0,F}$ farmers of genetic type N and $R_{0,F}$ farmers of genetic type X, because of assumption (ii), so the factor 2 does not appear before the number of such couples. Finally, although this is not necessary to perform the simulations, we can relate the population numbers at generation $t+1$ to those at the previous generation t by using equations (S7)-(S9) into equations (S11)-(S13), and the results into equations (S14)-(S16). This yield

$$P_{HG}(x,y,t+1) = R_{0,HG}[P_{HG}(x,y,t) - couples\ HX - couples\ HN], \quad \text{(S17)}$$

$$P_N(x,y,t+1) = R_{0,F}\ P_N(x,y,t)\,, \quad \text{(S18)}$$

$$P_X(x,y,t+1) = R_{0,F}[P_X(x,y,t) + couples\ HX + couples\ HN]. \quad \text{(S19)}$$

Besides this mathematical derivation of equations (S17)-(S19), it is also important to understand intuitively why, e.g., the number of couples *NX* does not appear in equations (S18)-(S19). The reason is that, although each *NX* couple implies that, e.g., one N individual less takes part in *NN* couples, i.e. that there are $R_{0,F}couples\ NX$ individuals of type N less the next generation, this is compensated by the fact that 50% of the $couples\ NX$ will also lead to individuals of type N (due to assumption (ii) above), thus contributing $0.5\big(2 \cdot R_{0,F}couples\ NH\big) = R_{0,F}couples\ NH$ individuals of type N to the next generation (remember that each couple has $2R_{0,F}$ children). For the same reason, NX couples do not appear in equation (S19), nor do HN couples appear in equation (S19). The latter do appear in equation (S19) because a HN couple does not imply that one X individual less takes part in *XX* couples, and thus its effect is not compensated. Couples *HX* do not appear in Eq. (S18) because all offspring of *HX* couples are farmers without haplogroup K, i.e. they all belong to group X (not to group N). They appear in Eq. (S19) because, although each *HX* couple implies that one X individual less takes part in *XX* couples (i.e., $R_{0,F}couples\ HX$ individuals less of type X the next generation), it also leads to $2R_{0,F}\ couples\ HX$ individuals of type X the next generation. We stress that equations (S17)-(S19) have been derived mathematically from equations (S11)-(S16), but we think that these explanations help to understand them intuitively.

If the number of individuals computed for some population group, cell, and time step is larger than its corresponding maximum ($P_{F\,max} = 3{,}200$ individuals/cell or $P_{HG\,max} = 160$ individuals/cell), then the simulation program sets it to the corresponding maximum value (this is applied, as in previous work [47, 139], to avoid population densities above saturation, which would not be biologically realistic). If $P_N + P_X > P_{F\,max}$, then P_N and P_X are both multiplied by $\frac{P_{F\,max}}{P_N+P_X}$, so that the new values satisfy that $P_N + P_X = P_{F\,max}$ and the proportion $\frac{P_N}{P_X}$ does not change. In equations (S14)-(S19), as in previous work [47, 139], we do not use a logistic growth function because it could lead to negative population numbers due to the fact that we are dealing with finite-difference equations (not with differential equations) [104, 139]. The solution of a logistic growth function (as applied in previous works [3, 183]) could be another alternative to avoid this problem, but we expect that it would yield similar results, so we do not apply it for mathematical simplicity.

5.8.6. Text S6. Estimation of the characteristic sea-travel distance from archaeological data

Previous work has shown the importance of long-distance sea travel in the spread of the Neolithic along the Mediterranean coast [47, 48, 245]. For this reason, our simulations include sea travel as a separate dispersal mechanism, in addition to inland travel. As in previous research by several authors [52, 139, 183, 51], we have estimated the characteristic distance of inland travel (50 km per generation) from ethnographic data for preindustrial farmers [52, 139]. Sufficiently detailed ethnographic data for sea travel distances of preindustrial farmers are unfortunately unavailable. In spite of this, we have estimated the characteristic distance of sea travel in the following way. Similarly to previous work [48, 246], we have required that the arrival times of the Neolithic at several regions along or near the Mediterranean (as predicted by our simulations) agree with that of the oldest archaeological data in each region, and that the spread routes correspond with those implied from archaeological data. In the simulations, sea travel takes place toward all coastal cells that can be reached in a straight line across the sea within a certain range.

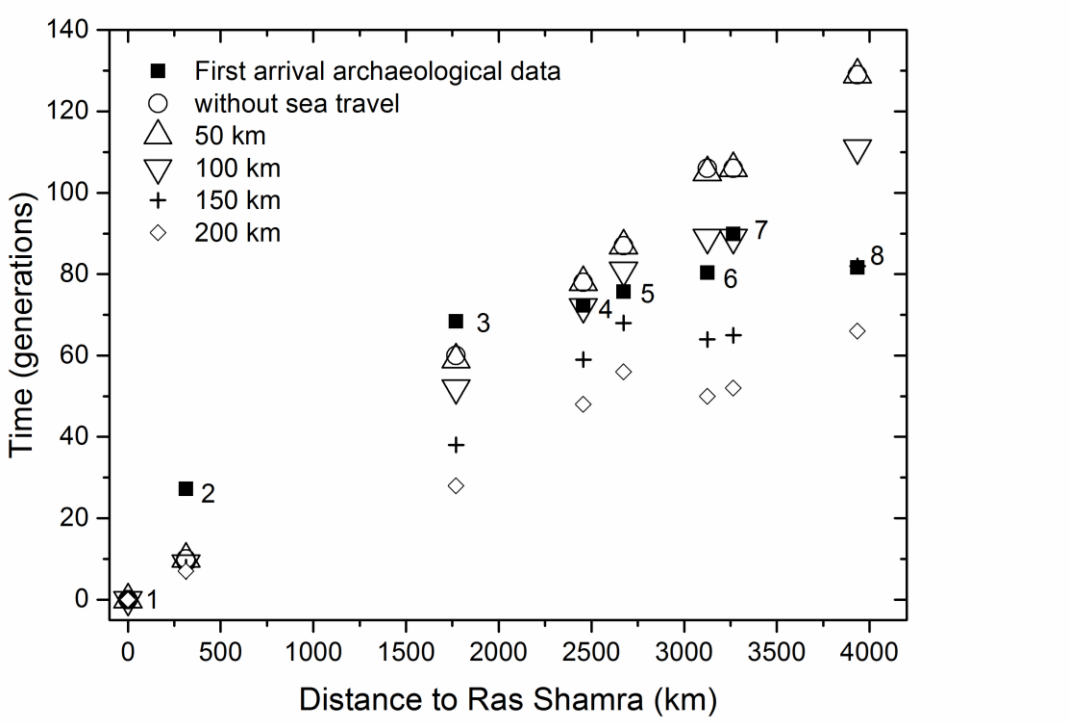


Figure 5.12 Estimation of the characteristic sea-travel range. Black squares: 1 oldest dates of the PPNB culture in Syria (Ras Shamra, 8,233 cal BCE [246], recall that PPNB is the Near-Eastern Neolithic culture that later spread into Europe [246]); earliest Neolithic dates in: 2 Anatolia (Hayaz Höyük, 7,361 cal BCE [246]), 3 Hungary-Croatia Starčevo (Gudnja, 6,044 cal BCE [246]), 4 Eastern Germany LBK (Dresden-Prohlis, 5,920 cal BCE [246]), 5 Western Germany LBK (Eilsleben, 5,811 cal BCE [246]), 6 North-Eastern Spain Cardial (Forcas, 5,661 cal BCE [247]), 7 Spain Navarre (Aizpea [at Basque Country], 5,357 cal BCE [247]), and 8 Portugal coastal Early Neolithic (Vale Pincel I, 5,620 cal BCE [247]). White symbols show the corresponding arrival times of our simulations with no sea travel (circles) and with sea travel of 50 km (up triangles), 100 km (down triangles), 150 km (crosses) and 200 km (rhombuses). The vertical axis is the time elapsed since the start of the simulations (8,233 BCE), measured in generations (1 generation = 32 yr [182]).

For the sake of clarity, we stress that the genetic data available (Appendix A Data S1) do not necessarily correspond to the earliest Neolithic sites in each region. The reason is that the genetic data, i.e. the individuals whose mtDNA haplogroup has been determined, have later (in some cases, substantially later) dates than those of the first Neolithic sites. Therefore, in order to compare to the arrival time obtained from our simulations, we cannot use the genetic dates. Instead, we have to use the observed arrival time of the Neolithic (i.e., the oldest archaeological data of Neolithic sites in the region considered). In Fig. 5.12, black squares correspond to the arrival dates of the Neolithic in the eight regions where we have the oldest genetic data. Note that in the main paper, Fig. 5.1, each square gives the time and distance of the oldest Neolithic *genetic* data in a region, whereas in Fig. 5.12 each square gives the time and distance of the oldest Neolithic *archaeological* site in that region (for this reason, the dates and distances in Fig. 5.1 and Fig. 5.12 are different). The information of the dates used in Fig. 5.12 is listed in its caption and in Appendix A Data S4.

We have performed our simulations with origin at Ras Shamra (oldest PPNB site in Syria) and different sea-travel ranges, assuming no population interaction. In Fig. 5.12 we show these results as white symbols, which correspond to our simulations with no sea travel (circles) and with sea travels up to 50 km (up triangles), 100 km (down triangles), 150 km (crosses) and 200 km (rhombuses). The arrival time of the Neolithic into a cell is recorded by the simulations as the generation when the farmer population of the cell reaches about a 10% of its maximum (this seems a reasonable percentage because it is unlikely that the archaeological record corresponds to the earliest farmers per region, and this values is close to the minimum size required for a human reproductive network to be viable [248]; however, changing this percentage would not change our conclusions). We can see in Fig. 5.12 (and in Appendix A Data S4) that apparently the best agreement between archaeological data (black squares) and the simulations is attained for sea travels up to 100 km (down triangles), since it provides a lower divergence between results. However, the results for this sea-travel range present two problems. (i) Simulations with sea travels up to 100 km arrive to regions 5, 6 and 8 later than the archaeological earliest data, which means that these results cannot really explain the earliest Neolithic evidences known, since the model arrives too late. (ii) A very important limitation of considering sea travels up to 100 km is that southern Italy is reached from the North (Fig. 5.13a), which is inconsistent with the archaeological dates that indicate very clearly that southern Italy was reached by sea from Albania or Greece (see Fig. 6 in reference [47]). In contrast, if we consider sea travels up to 150 km: (i) all regions are reached by the time of the earliest archeological date (see Fig. 5.12 and Appendix A Data S4). (ii) Crucially, southern Italy is appropriately reached before northern Italy through sea travel from Albania (see Fig. 5.13b). This is due simply to the fact that, in the simulation grid, the distance between the centers of the closest 50x50 km cells in Albania and Southern Italy is between 100 km and 150 km, so sea travels of at least 150 km are necessary for the front to enter Italy by this route. For reasons (i) and (ii) above, we consider that the best results are attained with sea travels of up to 150 km. It is interesting that the same result (i.e., 150 km) had been obtained previously by comparison to hundreds of individual sites (Table 1 and Fig. 8 in reference [47]). More detailed models, e.g. with a different sea travel distance in the Western [48] than in the Eastern Mediterranean could be considered, but we expect that they would not change our main result (namely, that the cline of haplogroup K implies that few farmers were involved in cultural diffusion).

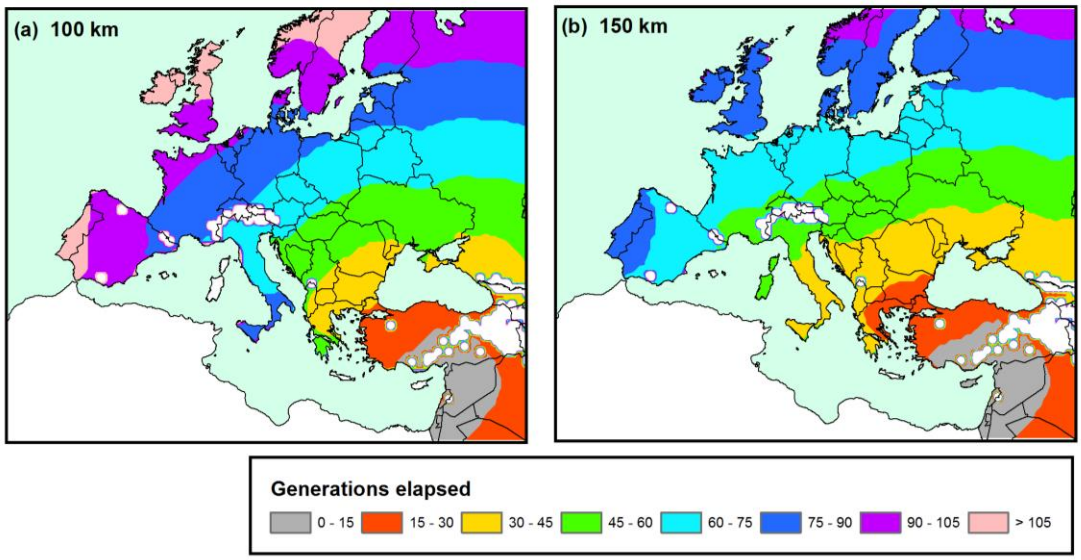


Figure 5.13 Predicted Neolithic arrival times computed with no interaction and for sea travel ranges of 100 km (a) and 150 km (b). White areas correspond to mountains, and the colors give the intervals of generations elapsed since the start of the simulations in Ras Shamra (Syria). Note that including interaction (cultural transmission) would not change the conclusion that the front enters Italy from the North in (a) and from the South in (b). Maps created with ArcMap 10 and the Spatial Analyst 10 extension (http://desktop.arcgis.com/es/desktop/).

5.8.7. Text S7. Implementation of the genetic initial conditions in the simulations

In order to compare the percentages of haplogroup K (%Ks) from our simulations to those from genetic data, we have to compute the %Ks from the simulations at the times of the genetic data (as given in the caption to Fig. 5.14), i.e. at the time when the fraction of ancient farmers bearing haplogroup K is known for each region (not at the time when the Neolithic arrived to it, which is obviously older and is given in the caption to Fig. 5.12).

As explained in the main paper (Sec. 5.4), we began our simulations at the date and location of the oldest Syrian PPNB site, namely Ras Shamra at 8,233 cal yr BCE [47]. Since this location is only about 150 km away from the average location of the Syrian sites with available mtDNA data, at first sight one might expect that we could directly apply the value (40%K) measured at the latter (Appendix A Data S2 and Data S3, estimated from the data reported by Fernández *et al.* [165]) also as initial genetic conditions at Ras Shamra. However, if we did so, we would obtain the results shown in Fig. 5.14. Note that in this figure the %K of PPNB Syrian sites (region 1) is not 40% but lower (except if $\eta = 0$). There are two reasons for this. The less important one is that the cell where we record the genetic information, located at the average location of the PPNB Syrian individuals in Appendix A Data S1, is 4 land-travel steps (50 km each) away from the origin of the simulation (i.e., the cell that contains Ras Shamra). Therefore, there is some interbreeding between the farmer population expanding from the original cell and the hunter-gatherer populations (which lack haplogroup K) at those other 4 cells.

However, the most important reason is that the simulation starts at 8,233 cal yr BCE (the date of Ras Shamra) but we compute the simulation results (lines in Fig. 5.14) for Syria (region 1) at 7,258 cal yr BCE (because 7,258 cal yr BCE is the average date of all PPNB individuals whose mtDNA haplogroup is known in this region, as computed in Appendix A Data S1 from the data in reference [165]). Therefore, the fact that the %K in region 1 in Fig. 5.14 is below 40% (except if $\eta = 0$) is mostly due to interbreeding between farmers and hunter-gatherers during the 1,000 yr elapsed since the beginning of the Neolithic (8,233 cal yr BCE) until the time when we have genetic data to compare to the simulations (7,258 cal yr BCE). Note that the decrease in %K in region 1 (Fig. 5.14) is larger the more intense the interbreeding (i.e. the higher the value of η), as it should.

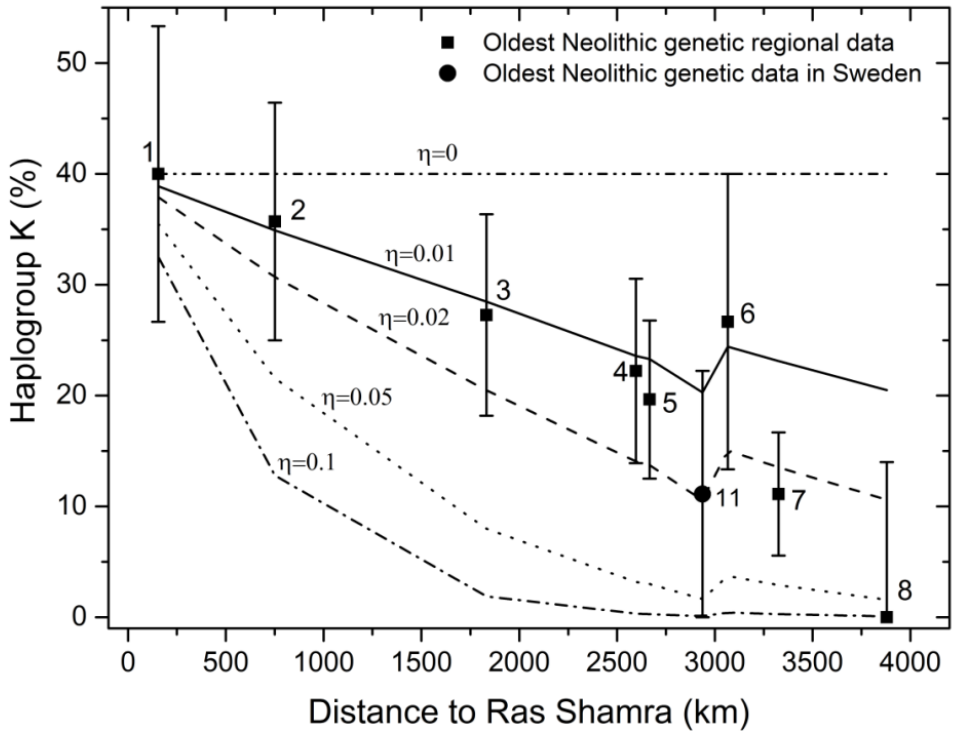


Figure 5.14 The lines are the model predictions when applying 40%K at the time of the oldest PPNB/C archaeological data in Syria (8,233 cal yr BCE). Symbols (with error bars) correspond to the observed percentages of haplogroup K in the 9 oldest regional cultures. Lines are the results from the simulations for different values of the interbreeding intensity η. The lines have been plotted by joining the simulation results for each of the 9 regional cultures (at its average location and date of its individuals). Here and in the rest of figures, for each regional culture, the date used to compute the results of the simulations is not that of the regional arrival of farming (as in Fig. 5.13) but the average date of the ancient individuals whose mtDNA haplogroup is known. In this way, we can compare simulated and observed %Ks. The regional cultures (and their average dates, as calculated in Appendix A Data S1) are: 1 Syria PPNB (7,258 cal yr BCE), 2 Anatolia (6,243 cal yr BCE), 3 Hungary-Croatia Starčevo (5,675 cal yr BCE), 4 Eastern Germany LBK (5,125 cal yr BCE), 5 Western Germany LBK (5,115 cal yr BCE), 6 North-Eastern Spain Cardial (5,286 cal yr BCE), 7 Spain Navarre (4,941 cal yr BCE), 8 Portugal coastal Early Neolithic (5,184 cal yr BCE) and 11 Sweden (2,802 cal yr BCE). The lines show the results of the simulations assuming that the %K in the Syrian region with PPNB sites was 40% at 8,233 cal yr BCE. However, according to the ancient DNA data available, this happened about 1,000 yr later (at 7,258 cal yr BCE). The problem is that in this figure, we do not obtain a 40% of haplogroup K in Syria at 7,258 cal yr BCE (see the values of the lines at region 1) except if $\eta = 0$ (no interbreeding and, therefore, no cline). In the main paper we applied a different implementation of the initial conditions to avoid this inconsistency (see Fig. 5.15).

In order to avoid this inconsistency, i.e. in order to avoid values of the percentage of haplogroup K below 40% at region 1 at time 7,258 cal yr BCE (lines in Fig. 5.14, region 1), we repeated the simulations by finding (by trial and error), for each value of η, an initial value (at time 8,233 cal yr BCE) for the %K in the starting cell (Ras Shamra) higher than 40% and such that the simulations yielded 40% of haplogroup K in region 1 at time 7,258 cal yr BCE (in agreement with the genetic data [165]). The results are shown in Fig. 5.15, which is the same as Fig. 5.3 in the main paper. Note that, as opposed to the results in Fig. 5.14, by taking into account the time lag between the first archaeological and genetic evidence, in Fig. 5.15 all lines predict a 40% of haplogroup K in region 1 (at 7,258 cal yr BCE), in agreement with the genetic data (Appendix A Data S1-Data S3) reported by Fernández *et al.* [165]. Therefore, in Fig. 5.15 the observed genetic initial condition (40%K in region 1, i.e. Syria) has been applied at the correct time (7,258 cal yr BCE). In contrast, in Fig. 5.14 the same genetic initial condition has been applied, but at an incorrect time (8,233 cal yr BCE).

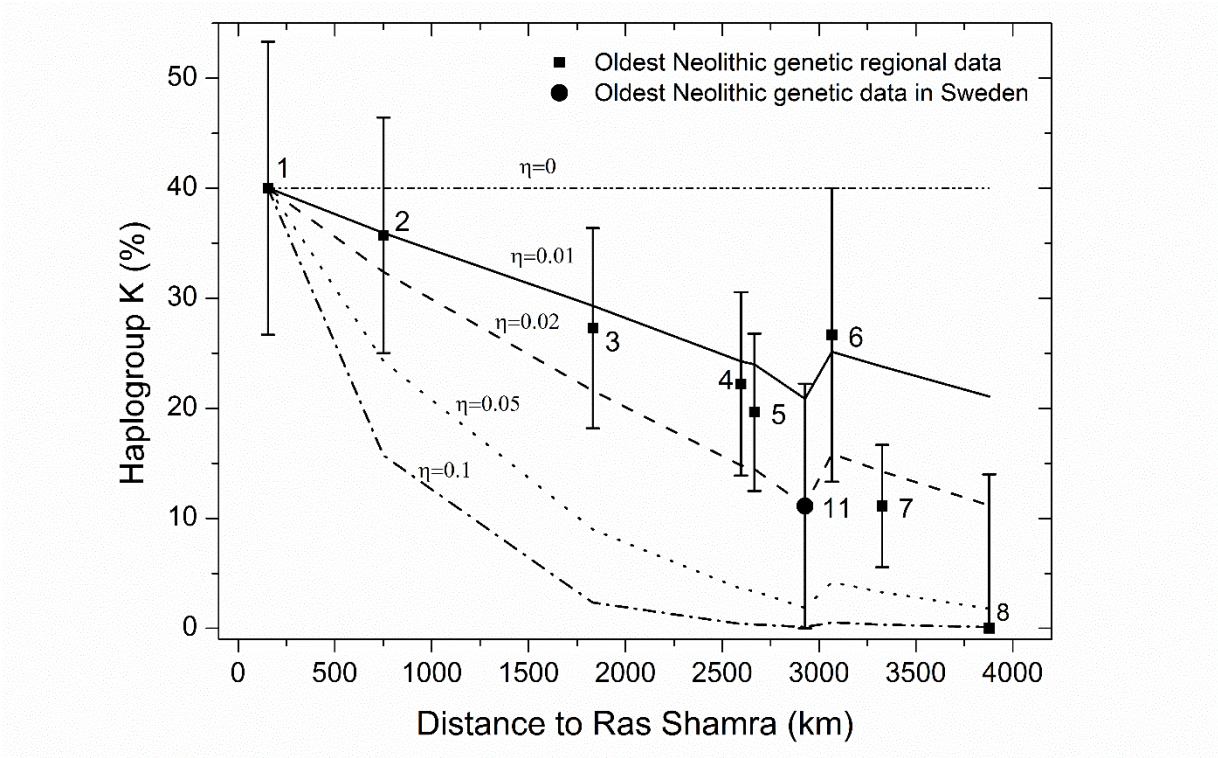


Figure 5.15 This figure is the same as Fig. 5.3 in the main paper. The lines are the model predictions when applying the adequate %K in Syria at 8,233 cal yr BCE to obtain a 40%K in Syria (region 1) at 7,258 cal yr BCE. Thus this figure shows the results of the simulations (lines) assuming that the percentage of haplogroup K in the Syrian region with PPNB sites was 40% at 7,258 yr BCE, in agreement with the ancient DNA data (symbol for region 1; percentage computed in Appendix A Data S1-Data S3 from the genetic data by Fernández *et al.* [165]). Compare to Fig. 5.14, where the %K at the initial cell is assumed to be 40% at 8,233 yr BCE instead. The regional cultures and dates are the same as in Fig. 5.14.

In all of our simulations, the maximum population density is 3,200 individuals/cell (see Sec. 5.8.5). Therefore, the initial genetic condition that at 7,258 cal yr BCE we had a 40%K in Syria (region 1) means that 1,280 of the 3,200 early farmers in this cell have haplogroup K. However, in the genetic dataset (Appendix A Data S1) we only have 6 of 15 individuals carrying K haplotypes, a value considerably

lower than in our simulations. Unfortunately, we have not been able to find further aDNA data for Early Neolithic individuals in Europe to improve our dataset (beyond the 15 individuals reported by Fernández *et al.* [165], already included in Appendix A Data S1). To check the representativeness of the dataset used, we have repeated our simulation but now using as initial genetic conditions the two extreme values (maximum and minimum) of the 80% CL error bar (note that the error bars have been computed, using the bootstrap method, precisely to take into account the small size of the available samples; see Sec. 5.8.10).

We have first repeated the computations in Fig. 5.15 (or Fig. 5.3) but using as initial %K in Syria (at 7,258 yr BCE) the lowest extreme of the error bar (the lowest extreme of the 80% CL range), i.e. 26.67%K, as shown in Fig. 5.16. Under these initial conditions we see that now the best fit between model and data is obtained for $\eta = 0.01$, since it is the value of the cultural transmission parameter for which the modelled results cross most of the error bars (in fact, all of them except '2 Anatolia'). Assuming a lower intensity of cultural transmission would yield a better prediction for Anatolia, but then the predictions would fall out of the measured range for the regions furthest from the origin. Therefore, if we initially had a 26.67%K in Syria (i.e., about 850 of the 3,200 early farmers), the observed cline could be explained assuming $\eta = 0.01$, an intensity of cultural transmission lower than the value $\eta = 0.02$ obtained when using the mean %K measured for Syria. This is as expected because a lower value of η leads to a smoother cline.

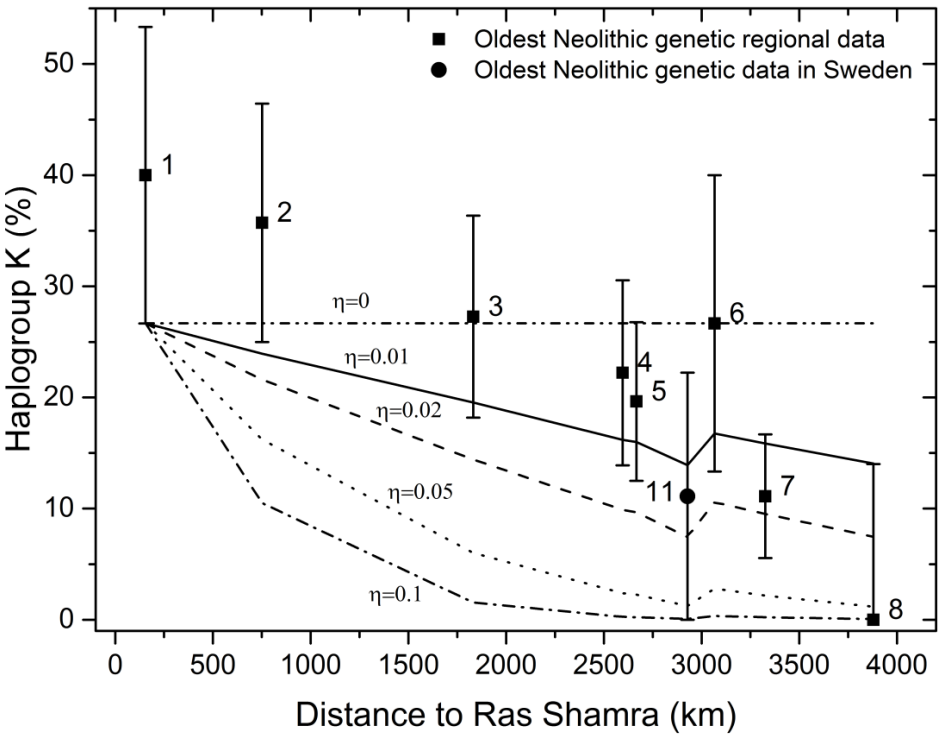


Figure 5.16 Model predictions when applying as initial genetic conditions in Syria (region 1) the *lower extreme* of the error bar of the observed %K (this is why the % in region 1 is not 40% but lower). This figure shows the results of the simulations (lines) assuming that the %K in the Syrian region with PPNB sites was 26,67% at 7,258 yr BCE, computed from the aDNA data as the lower extreme of the 80% CL bootstrap range (error bar for region 1; range computed in Appendix A Data S6 from the genetic data by Fernández *et al.* [165]). A good agreement with the data is obtained for $\eta \approx 0.01$. The regional cultures and dates are the same as in Fig. 5.14.

On the other hand, when we consider as initial genetic condition in Syria the upper extreme of the 80% CL range, i.e. 53.33%K, we see in Fig. 5.17 that the dashed line ($\eta = 0.02$) overestimates the percentage of the farmer population with haplogroup K at the regions furthest away from the origin. Thus, the intensity of cultural transmission needed to explain the cline is now higher than $\eta = 0.02$ (also as expected). The cline for $\eta = 0.03$, shown in Fig. 5.17 correctly predicts the observed percentages at the more distant populations(although it slightly underestimates the %K at regions 4 and 5). Therefore, the level of cultural transmission needed to explain the observed cline when assuming a 53.33%K in Syria at 7,258 BCE (i.e., about 1,700 of the 3,200 early farmers), is not higher than $\eta = 0.03$.

Thus, in summary, when considering the whole 80% CL range for the initial conditions, we have found that the observed genetic cline can be explained for intensities of cultural transmission in the range $\eta = 0.01 - 0.03$. Therefore, the conclusions in the main paper (that about 2% of farmers we involved in cultural transmission) is maintained (and refined by the range $2\% \pm 1\%$).

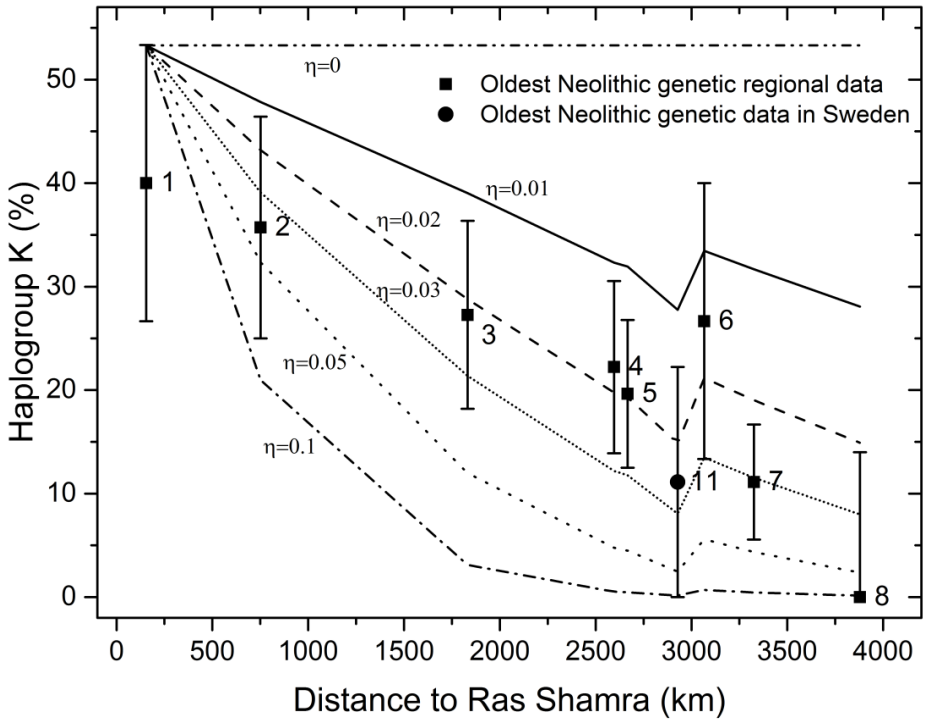


Figure 5.17 Model predictions when applying as initial genetic conditions in Syria (region 1) the *upper extreme* of the error bar of the observed %K (this is why the % in region 1 is not 40% but higher). This figure shows the results of the simulations (lines) assuming that the %K in the Syrian region with PPNB sites was 53,33% at 7,258 yr BCE, computed from the aDNA data as the upper extreme of the 80% CL bootstrap range (error bar for region 1; range computed in Appendix A Data S6 from the genetic data by Fernández *et al.* [165]). A good agreement with the data is obtained for $\eta \approx 0.03$. The regional cultures and dates are the same as in Fig. 5.14.

5.8.8. **Text S8. Understanding the minimum in the simulated clines**

In Fig. 5.15 (i.e., Fig. 5.3 in the main paper), we observe that the curves obtained from the simulations have a local minimum for region 11, i.e. Sweden. Interestingly, a minimum in Sweden is also seen for

the genetic data (squares and circle in Fig. 5.15). The general shape of the curves in Fig. 5.15 is easy to understand, as follows. As the distance (horizontal axis in Fig. 5.15) increases, we are considering regions further and further away from Syria (e.g., region 2 is Anatolia, region 3 is Hungary-Croatia, etc.). Since the time elapsed for the Neolithic front to reach a region tends to be larger the further away it is from Syria, there was more time for interbreeding between farmers and hunter-gatherers. This is why the percentage of haplogroup K (vertical axis in Fig. 5.15) tends to diminish with increasing distance (recall that hunter-gatherers lack haplogroup K).

We note in Fig. 5.15 (i.e., Fig. 5.3 in the main paper) that the tendency of decreasing percentage of haplogroup K with increasing distance from Syria is not always satisfied (there is a minimum in Sweden, region 11). The explanation of this subtle point is the following. As it is well-known, the Neolithic spread from Syria to Anatolia, then to Greece, and from there it followed two different routes. One was a Mediterranean route to Italy, France, Spain and Portugal. The other was a central/northern European route to Croatia, Germany, Denmark and Sweden [47, 95]. In order to see how this explains the minimum in Fig. 5.15, consider first a Neolithic front propagating along a coast. In this case, population dispersal can reach locations up to 150 km away (Sec. 5.8.6), measured in a straight line and across the sea (Sec. 5.4 in the main paper and Sec. 5.8.5). Now consider a Neolithic front propagating inland. In this case, dispersal is driven by jumps of about 50 km per generation (Sec. 5.4 and Sec. 5.8.5). Therefore, in order for the Neolithic front to travel a given distance, a *coastal* propagation obviously implies fewer jumps, i.e., fewer generations, and less time for interbreeding with hunter-gatherers than an *inland* propagation. Thus we should expect that a *coastal* propagation will lead, at a given distance, to a lower decrease of the percentage of haplogroup K (%K) than an *inland* propagation.

In Fig. 5.18 we have plotted the results of the simulations (for $\eta = 0.02$) for the two routes mentioned above (i.e., the Mediterranean and the central/northern European ones) separately. Up to Greece, both routes are the same and thus lead to the same %K as a function of distance. However, after Greece the Mediterranean route is mostly coastal (to France, Spain and Portugal), in sharp contrast with the central/northern European route, which is mostly inland (to Germany, Denmark and Sweden). Thus the %K of the central/northern European route becomes smaller than that of the Mediterranean route, for the reasons argued in the previous paragraph (see the slope change in the central/northern European route after its coastal spread ends up in Trieste in Fig. 5.18).

Now that we have understood the shape of the curve for each route (Fig. 5.18), we can explain the minimum in Fig. 5.15, as follows. If in Fig. 5.18 we joined the three points Germany-Sweden-NE Spain, we would obtain a minimum. This is precisely the minimum in Fig. 5.15 (where regions 4-5 are again Germany, region 11 is Sweden, and region 6 is NE Spain). Thus, the minimum in Fig. 5.15 is due to the existence of two propagation routes for the European Neolithic. These are the Mediterranean and the central/northern European routes, which are respectively (for large distances) a coastal route (with high %Ks) and an inland route (with lower %Ks), as seen in Fig. 5.18. Hence, the minimum in Fig. 5.15 is a purely geographical effect, due to the presence of the Mediterranean Sea. Below we will check this last point in another way (namely, by showing that simulations without sea display no minimum). However, before doing so, it is important to consider several related issues.

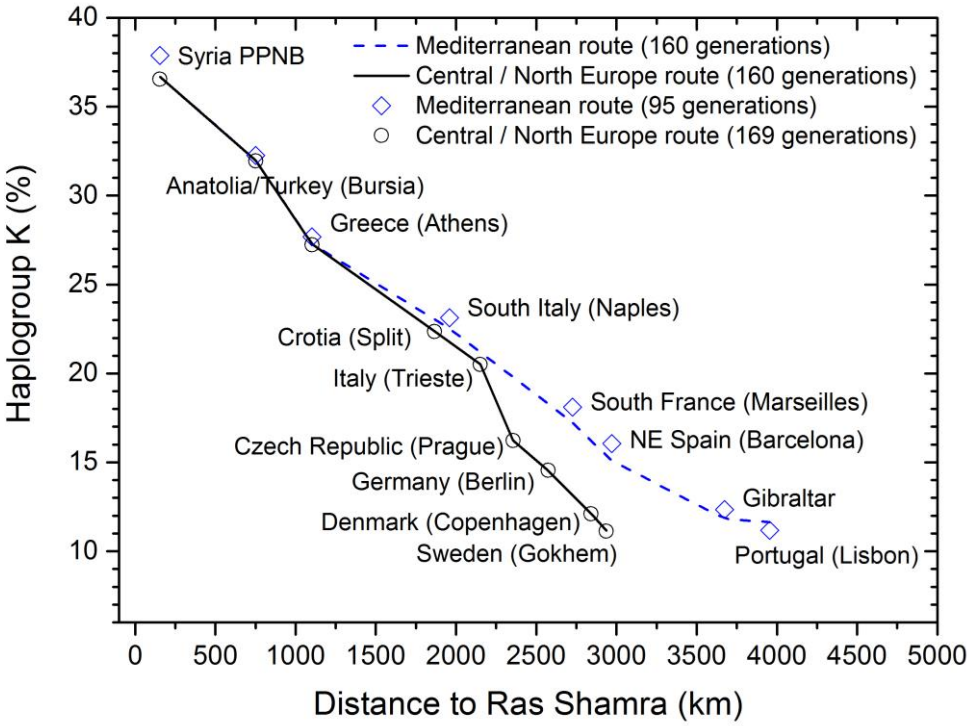


Figure 5.18 Results of the simulations for $\eta = 0.02$ along two spread routes. This figure is similar to, e.g., Fig. 5.15 (i.e., Fig. 5.3 in the main paper), but instead of considering the regions for which the DNA of ancient farmers has been determined, here we consider several locations on the two main routes along which the Neolithic spread, namely the Mediterranean route (dashed blue line) and the central/northern European route (solid black line). Another difference with Fig. 5.15 is that both lines in this figure have been obtained from our simulations at a single time, namely 3,113 yr BCE (i.e., 160 generations after the departure of the wave of advance from Ras Shamra, Syria). We have chosen this time so that the Neolithic wave of advance has reached all of Europe. However, in order to make sure that this figure can be used to understand the minimum in Fig. 5.15 (in spite of having used a value of time different from those used in Fig. 5.15), we also include the following results at other times. Blue rhombuses are results at locations on the Mediterranean route obtained at the most recent genetic date on that route, i.e. at the time of the genetic data of Portugal in Fig. 5.15 (5,184 yr BCE or 95 generations). Empty circles are the results at locations on the central/northern European route obtained at the most recent genetic date on that route, i.e. at the time of the genetic data of Sweden in Fig. 5.15 (2,802 yr BCE or 169 generations). We can see that considering different times leads to almost the same results, so the explanation of the minimum in Fig. 5.15 remains valid.

Genetic data from modern populations display distinct clines along the Mediterranean and central/northern European directions, and it has been suggested that this difference may be due to the respective routes of Neolithic dispersal [60]. Unfortunately, ancient mtDNA *data* are not yet numerous enough to distinguish whether both routes led to distinct ancient clines of haplogroup K or not. However, the presence of the minimum in Sweden, both according to the model *simulations* and to the data available (squares and circle in Fig. 5.15, i.e. Fig. 5.3 in the main paper), strongly suggests this possibility (see Fig. 5.18). Nevertheless, we cannot yet plot both *observed* clines (in contrast to the *simulated* ones in Fig. 5.18) due to the paucity of aDNA data available at present. Indeed, in Italy there are no mtDNA data from ancient farmers yet. In Greece, the mtDNA of only one early Neolithic individual is known [171]. In many other regions, there are data only from a small number of individuals (Fig. 5.2).

In Fig. 5.18, at the beginning of the spread, e.g. in Syria, the %K of the rhombuses is higher than that of the circles. This is reasonable, because rhombuses correspond to the %K at an earlier time than circles, thus less interbreeding with HGs has taken place. Note that this decrease (from the rhombus to the circle) is lower in Anatolia or Greece than in Syria, because populations of farmers with higher %K than the local frequency arrive to Anatolia and Greece but not to Syria (the origin of the expansion). On the other hand, Portugal is the only region where the %K increases with time (rhombus and dotted line in Fig. 5.18), because populations with higher %K arrive from the North, the South and the East into Portugal (see Fig. 5.13b).

Sweden is the latest region in Europe where the Neolithic arrived, and therefore the region with most recent DNA data in Fig. 5.15. Since more time implies more interbreeding with hunter-gatherers, and therefore a larger decrease in the %K, we could expect that the presence of the minimum in Sweden in Fig. 5.15 is due to the fact that we have plotted the %K as a function of distance, not of time. In order to check this point, in Fig. 5.19 we plot the %K as a function of time (not of distance as in Fig. 5.15). We observe that a minimum still appears. However, now the minimum does not correspond to Sweden (as in Fig. 5.15) but to Portugal (Fig. 5.19).

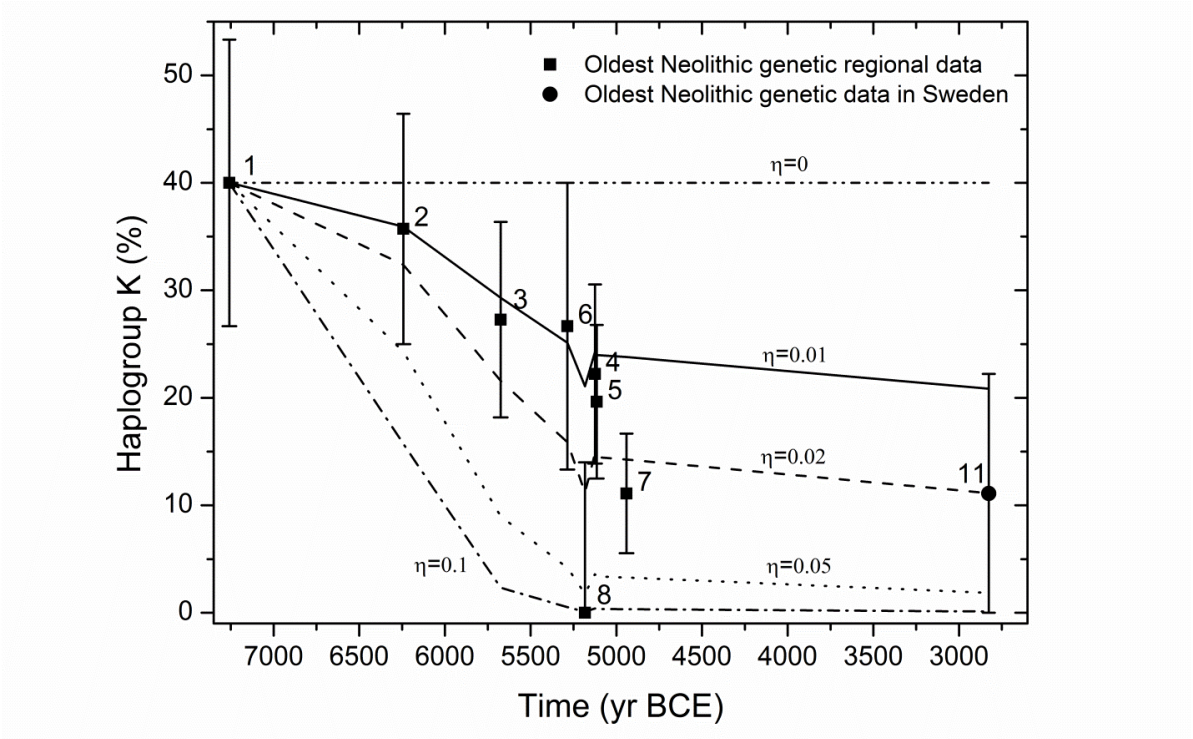


Figure 5.19 Percentage of mtDNA haplogroup K, *as a function of time*. The data points correspond to the same 9 regional cultures which have been plotted *as a function of distance* in Fig. 5.15, namely: 1 Syria PPNB, 2 Anatolia, 3 Hungary-Croatia Starčevo, 4 Eastern Germany LBK, 5 Western Germany LBK, 6 North-Eastern Spain Cardial, 7 Spain Navarre, 8 Portugal coastal Early Neolithic and 11 Sweden. The error bars are the 80% CL for the %K (i.e., the same as in, e.g., Fig. 5.15 or Fig. 5.3 in the main paper). The lines join the results of the simulations for different values of the cultural diffusion intensity η. Note that, in contrast to Fig. 5.15, region 11 (Sweden) appears at the right-hand side, and region 8 (Portugal) in the middle of the plot, where the minimum is now located.

We can explain the minimum in Fig. 5.19, again in terms of the Mediterranean and central/northern European routes, as follows. In Fig. 5.20 we plot the results of the simulations (for $\eta = 0.02$) for the

two routes separately but as a function of time (not of distance as in Fig. 5.18). Similarly to our explanation above of the minimum in Fig. 5.15 (from the two routes in Fig. 5.18), we note that a minimum would appear in Portugal (region 8) in Fig. 5.20 if we joined regions 6-8-4/5. This is precisely the reason why we now see a minimum in Portugal (region 8) in Fig. 5.19 (rather than in Sweden as in Fig. 5.15).

The fact that the minimum appears in Portugal if time is the horizontal axis (Figs. 5.19-5.20) whereas it appears in Sweden if distance is the horizontal axis (Fig. 5.15) can be understood as follows. Note that the value of the vertical axis (simulated %K) for each regional culture is the same in Figs. 5.19-5.20 as for the dashed line in Fig. 5.15. Regional cultures 4, 5 (Germany) and 11 (Sweden) appear to the left of culture 8 (Portugal) in Fig. 5.15 because their distances are lower (so the minimum is in Sweden). However, they appear to the right of culture 8 (Portugal) in Figs. 5.19-5.20 because their dates are later (so the minimum is in Portugal), simply because the average dates of the ancient individuals whose mtDNA haplogroup is known are more recent for cultures 4, 5 and 11 than for 8.

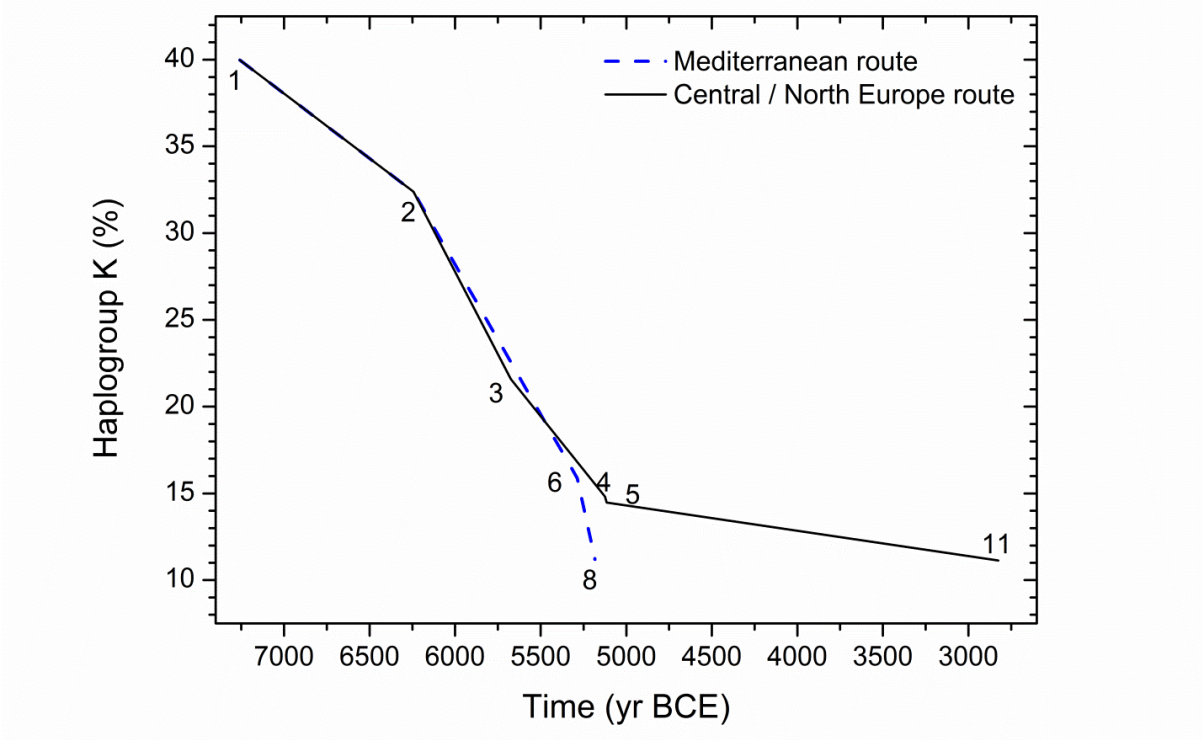


Figure 5.20 Results of the simulations for $\eta = 0.02$, as a function of time, for the regional cultures in Fig. 5.19 located on the Mediterranean route (dashed blue line) and the central/northern European route (solid black line). In contrast to Fig. 5.18, here we cannot consider a single value of time (because here the horizontal axis is time, not distance). Thus we consider the same regions and their values of time as in Fig. 5.19 (the value of time in each region being equal to the average date of the individuals whose mtDNA is known). A minimum would appear in Portugal (region 8) if we joined regions 6-8-4/5. These are precisely the regions where the minimum also appears in Fig. 5.19. This explains the minimum in Fig. 5.19.

Finally, we checked in another way that the minimum of the percentage of haplogroup K is indeed due to the geography of Europe. We simulated the spread of the Neolithic and its genetic dynamics using, instead of a map in Europe, a homogeneous space, i.e., a grid with only land nodes (without any

seas neither mountains), so that all individuals that change their residence move 50 km (this simulation can be performed by modifying only the grid and initial conditions in Program S1, but for convenience we have made all of the necessary files available at the journal web as Program S4, or at http://copernic.udg.es/QuimFort/2017_08_07r__Program_S4.zip). We run our simulations on a grid of the same size as the geographically realistic grid (180x102 cells of 50kmx50km each). For simplicity, we have set the origin of the spread at the same node and with the same initial genetic conditions as in Fig. 5.15 (for $\eta = 0.02$) and Figs. 5.18-5.20, namely, the node containing Ras Shamra, which has coordinates (112, 31) and initially (8,233 cal yr BCE) a percentage of farmers with haplogroup K equal to 42.2%K, which is the percentage needed in the previous simulations (in real geography) so that the 40%K in 1 Syria is correctly predicted at 7,258 cal yr BCE when $\eta = 0.02$ (see Fig. 5.15 and Figs. 5.19-5.20). In Fig. 5.21 we show the results of our simulations for $\eta = 0.02$ at the same cells used in the previous figures (e.g. Fig. 5.15), which we have labelled accordingly. However, since we are now dealing with homogeneous space, the results do not really correspond to the regional cultures in the previous figures (e.g., "1 Syria PPNB" or "11 Sweden"), but to points located at the same radial distance from the origin as them, for which we have computed the %K at 200, 500 and 900 generations after the beginning of the spread (Fig. 5.21). For clarity we mention that in the node corresponding to the average location of region '1. Syria PPNB', i.e. node (115, 32), we find at 7,258 cal yr BCE a percentage equal to exactly 39.98% in real geographies (Fig. 5.15 for $\eta = 0.02$ and Figs. 5.19-5.20) and 39.91% in homogeneous space (i.e., about 40% in both cases). However, in Fig. 5.21 (homogeneous space) we obtain for the upper line about 36% (rather than 40%) because this result is after 200 generations (whereas the percentage 40% is obtained at 7,258 cal yr BCE, i.e. 30 generations after 8,233 cal yr BCE). For a Neolithic wave spreading in homogeneous space, we simply expect that the percentage of haplogroup K will diminish with increasing distance, and that this cline will gradually disappear as time passes (both features being due to interbreeding). This is precisely what can be observed from our simulations in Fig. 5.21, but most importantly, in contrast with Fig. 5.15, there is no local minimum in Fig. 5.21. Thus the minimum in, e.g., Fig. 5.15 indeed arises due to the presence of the Mediterranean sea in Europe, which leads to the existence of two expansion routes with differentiated dispersal behavior, namely the central/northern European route (which is mainly inland and has thus jumps of 50 km per generation) and the Mediterranean one (which is mainly coastal and has thus jumps of up to 150 km per generation).

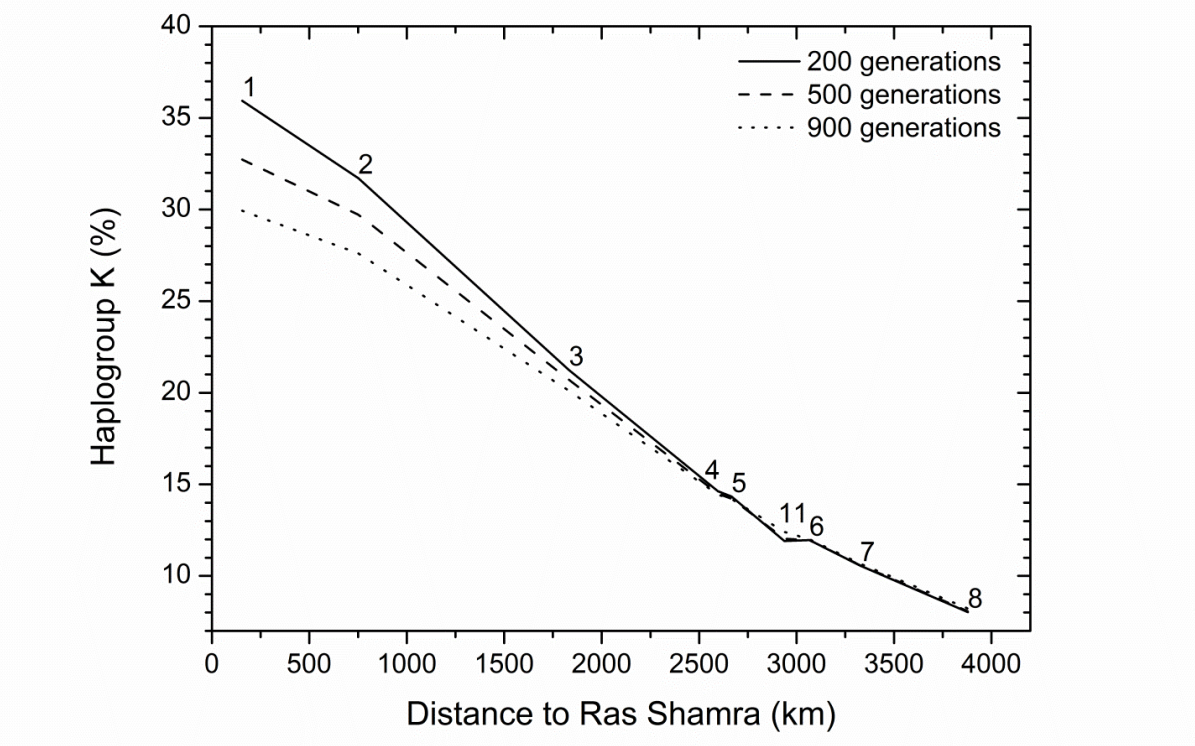


Figure 5.21 This figure shows the results of the simulations (for $\eta = 0.02$) without seas neither mountains in the simulation grid. We have set the start of the spread at the same cell and initial genetic conditions as in Fig. 5.15 and Figs. 5.18-5.20, and the simulated results plotted correspond to the same cells as the cultural regions in Figs. 5.15-5.17 and Figs. 5.19-5.20 (labelled accordingly in the figure) but at 200, 500 and 900 generations after the origin of the spread. As expected, the %K decreases with increasing distance from the origin of the spread, and the cline is gradually erased with time (1 generation=32 yr [182]). In contrast to the line in Fig. 5.15 for $\eta = 0.02$, no minimum appears here (homogeneous space). This confirms that the minimum in, e.g., Fig. 5.15 (which is the same as Fig. 5.3 in the main paper) is a purely geographical effect, due to the existence of the Mediterranean sea.

5.8.9. **Text S9. Horizontal/oblique transmission**

All models in the main paper and other sections in this Supplementary Information use the equations of vertical transmission, i.e. interbreeding between farmers and hunter-gatherers. In this section we show that the conclusions would not change if we considered, instead, acculturation, i.e. learning of agriculture by hunter-gatherers from farmers of the same generation (horizontal transmission) and/or the previous one (oblique transmission).

Vertical transmission leads to the following new population numbers (in each spatial cell) after one generation (see equations (S17)-(S19)),

$$P_{HG}(x,y,t+1) = R_{0,HG}[P_{HG}(x,y,t) - couples\ HX - couples\ HN], \quad \text{(S20)}$$

$$P_N(x,y,t+1) = R_{0,F}\ P_N(x,y,t)\,, \quad \text{(S21)}$$

$$P_X(x,y,t+1) = R_{0,F}[P_X(x,y,t) + couples\ HX + couples\ HN], \quad \text{(S22)}$$

where, for the cell considered, P_{HG} is the number of hunter-gatherers, P_N is the number of farmers who have haplogroup K, and P_X is the number of farmers who do not have haplogroup K. The numbers of mixed couples are given by equations (S5)-(S6), namely

$$couples\ HN = \eta \frac{P_{HG}(x,y,t) \cdot P_N(x,y,t)}{P_{HG}(x,y,t) + P_F(x,y,t)}, \tag{S23}$$

$$couples\ HX = \eta \frac{P_{HG}(x,y,t) \cdot P_X(x,y,t)}{P_{HG}(x,y,t) + P_F(x,y,t)}, \tag{S24}$$

and the total number of farmers in the spatial cell considered is $P_F(x,y,t) = P_N(x,y,t) + P_X(x,y,t)$.

We can interpret the meaning of η by noting that for pioneering, low-density populations of farmers ($P_N \approx 0$, $P_X \approx 0$ and thus $P_F \approx 0$), equations (S21)-(S24) for the special case $R_{0,F} = 1$ (no net reproduction) lead to $P_F(x,y,t+1) - P_F(x,y,t) \approx \eta P_F(x,y,t)$, so that η can be interpreted as the relative increase in the number of farmers per generation due to interbreeding with HGs (i.e., the proportion of farmers that take part in vertical cultural transmission).

For horizontal/oblique transmission [3], the first three equations are valid replacing the number of couples by the number of hunter-gatherers who learn farming (which we call converts and do not have haplogroup K, i.e. they belong to population X) from each population of farmers (X and N), i.e.

$$P_{HG}(x,y,t+1) = R_{0,HG}[P_{HG}(x,y,t) - converts\ HX - converts\ HN], \tag{S25}$$

$$P_N(x,y,t+1) = R_{0,F}\ P_N(x,y,t)\ , \tag{S26}$$

$$P_X(x,y,t+1) = R_{0,F}[P_X(x,y,t) + converts\ HX + converts\ HN], \tag{S27}$$

where [3]

$$converts\ HN = f \frac{P_{HG}(x,y,t) \cdot P_N(x,y,t)}{\gamma P_{HG}(x,y,t) + P_F(x,y,t)}, \tag{S28}$$

$$converts\ HX = f \frac{P_{HG}(x,y,t) \cdot P_X(x,y,t)}{\gamma P_{HG}(x,y,t) + P_F(x,y,t)}. \tag{S29}$$

Analogously to the paragraph below Eq. (S24), we can interpret the meaning of C by noting that for pioneering populations of farmers ($P_N \approx 0$, $P_X \approx 0$ and thus $P_F \approx 0$), equations (S26)-(S29) for the special case $R_{0,F} = 1$ (no net reproduction) lead to $P_F(x,y,t+1) - P_F(x,y,t) \approx C P_F(x,y,t)$, so that $C \equiv f/\gamma$ can be interpreted as the relative increase in the number of farmers per generation due to acculturation with HGs (which is the same, if $C < 1$, as the proportion of farmers that take part in horizontal/oblique cultural transmission) [3].

In the simple case $\gamma = 1$ (which corresponds to random copying of behavior between individuals [3]), it is easy to see that $0 \leq f \leq 1$ (otherwise, $P_{HG}(x,y,t+1)$ could become negative for $P_{HG} << P_F$). Then equations (S25)-(S29) for horizontal/oblique transmission are the same as equations (S20)-(S24) for vertical transmission, with η replaced by f. Recall also that for vertical transmission $0 \leq \eta \leq 1$ [94]. Thus, the same model as in the main paper can be used for horizontal/oblique transmission, instead of vertical transmission. Obviously, for horizontal/oblique transmission the conclusion (from

Fig. 5.3, or Fig. 5.15) would be that $f = 0.02$ (instead of $\eta = 0.02$), i.e. that about 2% of new farmers join the pioneering farming populations per generation due to acculturation of hunter-gatherers (instead of due to interbreeding with hunter-gatherers) or, equivalently, that about 2% of farmers teach agriculture to a hunter-gatherer (instead of mating a hunter-gatherers).

A more general case is to consider both horizontal/oblique transmission (acculturation) and vertical transmission (interbreeding). In such an instance, the corresponding equations (as given above) should be applied sequentially in the simulations, in general with different values for parameter f (horizontal/oblique transmission) and η (vertical transmission). Accordingly, the equations are a bit more complicated, because vertical transmission makes the frequencies of parent and children different, so it must be taken into account explicitly that the teachers belong to the parental generation in oblique transmission but not in horizontal transmission (compare Eqs. (3.4.1) to (3.1.3) in Ref. [93]). We do not perform such simulations, for the following reason. We have estimated the value of η (namely $\eta \approx 0.02$) in Fig. 5.3 in the main paper, by assuming only vertical transmission. Alternatively, if we considered only horizontal/oblique transmission, we would estimate the same value for f (i.e., $f \approx 0.02$). But if we considered a model with both kinds of transmission, we would have at least two independent parameters (η and f), and we cannot estimate both of them univocally from the genetic data available (i.e., from the error bars in Fig. 5.3 in the main paper). However, we next show that the conclusions in the main paper would not change under such more complicated models. Clearly, values $\eta \approx 0.02$ and $f \approx 0.02$ or higher would yield more cultural transmission than the case considered in the main paper (Fig. 5.3), i.e. $\eta \approx 0.02$ and $f = 0$. Therefore, for values $\eta \approx 0.02$ and $f \approx 0.02$ or higher, obviously the simulated cline will be too steep to be consistent with the genetic data (error bars in Fig. 5.3). Thus, we can assure that more complicated models (i.e., with both vertical and horizontal/oblique transmission) will be consistent with the genetic data only if $\eta < 0.02$ and $f < 0.02$ (more precisely, we should expect e.g. $\eta + f \leq 0.02$). These values are very small compared to the maximum possible ones (namely $\eta = 1$ and $f = 1$). Noting that, in regions where the first famers arrived ($P_N \approx 0$ and $P_X \approx 0$), the equations above simplify to $couples\ HN \approx \eta P_N$, $couples\ HX \approx \eta P_X$ (for vertical transmission) and $converts\ HN \approx f P_N$, $converts\ HN \approx f P_X$ (for horizontal/oblique transmission), we can interpret the result above ($\eta + f < 0.02$) by stating that less than 2% of farmers took part in cultural transmission, either by mating with hunter-gatherers or by teaching agriculture to them. Thus, about 98% of the population did not take part in cultural transmission. In this sense, cultural diffusion was of little importance. Therefore, the main conclusion of our work remains the same, regardless that we consider vertical, horizontal, oblique, or any combination of these three kinds of cultural transmission.

5.8.10. Text S10. Calculation of the error bars of percentages of haplogroup K

For each sample (e.g. Syria PPNB, Anatolia, Hungary-Croatia Starčevo, etc.), we calculated the 80% confidence-level (CL) range of its percentage of haplogroup K (hereafter called %K), which we represent as error bars in Figs. 2-3, by bootstrap case resampling. In order to do so, we drew 10,000 random resamples from each original sample with replacement. Each resample had the same number of individuals as the original sample (e.g., 15 individuals for Syria PPNB, 28 for Anatolia, etc.). For these 10,000 resamples, we computed a histogram with the number of resamples versus their %K. Then, with 80% CL, the error bar is limited by the 10% and 90% quartiles of this distribution (i.e. the values

of the %K below which there are 10% and 90% of the histogram resamples, respectively). We performed these calculations using Mathematica and checked them using Excel.

However, this bootstrap method cannot be applied to the case of populations with 0%K (e.g., Portugal coastal Early Neolithic), simply because then there are no individuals with haplogroup K, so all bootstrap samples have 0 individuals with haplogroup K. This would yield a vanishing error bar for the estimation 0%K, which is not reasonable, for the following reason. For example, for the sites in the sample called 'Portugal coastal Early Neolithic' there are only 10 individuals. None of them has haplogroup K, so its frequency is obviously 0%. However, if we had e.g. 100 individuals, and none of them had haplogroup K, its frequency would again be 0% but with more certainty, i.e., the error bar should be narrower. Thus, assigning a vanishing error bar to samples with 0% of a haplogroup is not justified. In order to deal with such samples, we could begin by introducing reasonable assumptions, e.g. by adding noise to the data [249, 250, 251]. However, such approaches would require hypotheses (on the kind of noise, its parameter values, etc.) [249, 250, 251]. Clearly, it would be better to find a solution without introducing such assumptions. With this aim, we devised the following method.

Although our method is general, for clarity let us consider a specific sample we are interested in, e.g. 'Portugal coastal Early Neolithic'. As mentioned above, in this sample there are only 10 individuals and none of them carried haplotypes from haplogroup K, so the K frequency is 0%. For the sake of simplicity, as a first step, imagine 11 possible populations (each of them composed of a very large number of individuals), with percentages of individuals with haplogroups different than K (which we call "0" individuals) equal to 100%, 90%, 80%, ..., 20%, 10% and 0%. We call those populations P_{100}, P_{90}, ..., P_{10}, P_{0}, respectively (note that they have %s of the K haplogroup equal to 0%, 10%, ..., 90% and 100%, respectively). Imagine that we choose at random 1 of these 11 populations, next we choose 10 individuals at random from it, and it turns out that all of them are "0" individuals (i.e., none of them has haplogroup K, as in the case of 'Portugal coastal Early Neolithic'). In such a situation, obviously it is more likely that we have chosen the population P_{100} than P_{90}, it is also more likely that we have chosen the population P_{90} than P_{80}, etc. But what are the exact probabilities that we have chosen each population? The probability of population P_{100} for the situation considered (i.e., that in which all 10 individuals are "0") is

$$p(\mathrm{P}_{100} \mid 0000000000) = \frac{p(\mathrm{P}_{100} \cap\, 0000000000)}{p(0000000000)}, \tag{S30}$$

where the symbol ∩ denotes intersection, i.e. co-occurrence of the two events, and

$$p(\mathrm{P}_{100} \cap\ 0000000000) = \frac{number\ of\ cases\ \mathrm{P}_{100}\ and\ 0000000000}{number\ of\ total\ cases}, \tag{S31}$$

$$p(0000000000) = \frac{number\ of\ cases\ 0000000000}{number\ of\ total\ cases}\ , \tag{S32}$$

and the number of total cases includes all 11 possible populations and all possible outcomes besides 0000000000 (e.g., 1000000000, 010000000, 1100000000, etc.). Similarly for the other populations,

$$p(\mathrm{P}_{90} \mid 0000000000) = \frac{p(\mathrm{P}_{90} \cap\, 0000000000)}{p(0000000000)}, \tag{S33}$$

$$p(\mathrm{P}_{80} \mid 0000000000) = \frac{p(\mathrm{P}_{80} \cap\, 0000000000)}{p(0000000000)}, \tag{S34}$$

etc. Clearly, since they refer to two independent events, the numerators in equations (S30), (S33), (S34), etc. are equal to the probability that we have chosen the considered population P_i (namely $\frac{1}{11}$, because it has been chosen at random) times the probability that, if we have chosen this population, we have also chosen a sample in which all 10 individuals are "0". Thus

$$p(\mathrm{P}_{100} \cap\ 0000000000) = \frac{1}{11}1 = \frac{1}{11}, \tag{S35}$$

$$p(\mathrm{P}_{90} \cap\ 0000000000) = \frac{1}{11}(0.9^{10}) = \frac{0.9^{10}}{11}, \tag{S36}$$

$$p(\mathrm{P}_{80} \cap\ 0000000000) = \frac{1}{11}(0.8^{10}) = \frac{0.8^{10}}{11}, \tag{S37}$$

etc. By adding up these values, equation (S32) can be written as

$$p(0000000000) = \frac{1}{11}(1 + 0.9^{10} + 0.8^{10} + \cdots + 0.1^{10} + 0^{10}), \tag{S38}$$

and we find the final result for population P_{100} from equations (S30), (S35) and (S38) as

$$p(\mathrm{P}_{100} \mid 0000000000) = \frac{1}{1+0.9^{10}+0.8^{10}+\cdots+0.1^{10}} = 0.6705. \tag{S39}$$

Similarly we find, for the other populations,

$$p(\mathrm{P}_{90} \mid 0000000000) = \frac{0.9^{10}}{1+0.9^{10}+0.8^{10}+\cdots+0.1^{10}} = 0.2338, \tag{S40}$$

$$p(\mathrm{P}_{80} \mid 0000000000) = \frac{0.8^{10}}{1+0.9^{10}+0.8^{10}+\cdots+0.1^{10}} = 0.0720, \tag{S41}$$

etc. As expected, the probability is highest for population P_{100} (i.e., for 0%K). The important point is that, in contrast to the bootstrap method with case resampling (which would predict that P_{100} is the only possible population), we have computed non-vanishing probabilities for the other populations (and they decrease with increasing %K, also as expected). Moreover, by adding equations (S39) and (S40), we find that

$$p(\mathrm{P}_{100} \mid 0000000000) + p(\mathrm{P}_{90} \mid 0000000000) = 0.9043, \tag{S42}$$

from which we can state that there is a probability of 90.43% that our sample comes from a population with a percentage of "0" individuals between 100% and 90%. In other words, we have found, with 90.43% confidence level, that our sample comes from a population in which the percentage of haplogroup K is in the range 0%-10%K. Note, however, that in this first computation (i.e., using 11 possible populations) there is a lot of uncertainty, because the closest possible result that we can possibly estimate would be obtained by adding $p(\mathrm{P}_{80} \mid 0000000000)$ to equation (S42) and then, the range of the percentage of "0" individuals would be between 100% and 80%, i.e., the upper limit of of the % of haplogroup K would be 20%K, rather than 10%K as above. Thus, it is safe to accept that there is an error of up to 10% in the estimation of percentages using 11 populations (another way to see this is simply to note that our 11 possible populations are separated by increases of 10%K). Therefore, our conclusion should be that, with 90.43% confidence level, the original population had a frequency in the range 0%-20%K.

Note also that, for the CL we are interested in, namely 80% (because this is the range used in the main paper), this first calculation does not lead to a precise range of the %K, because we can only estimate such a range with a 67.05% CL (using equation (S39)), with a 90.43% CL (using equation (S42)), etc. We next show that we can solve this problem by considering a larger number of possible populations.

Secondly, we repeat the previous procedure with 101 populations (instead of 11 as above), with percentages of "0" individuals equal to 100%, 99%, 98%, ..., 2%, 1% and 0 %. Thirdly, we repeat the same approach with 1,001 populations (with percentages 100%, 99.9%, ..., 0.1% and 0%). And fourthly, we do the same with 10,001 populations (with percentages 100%, 99.99%, ..., 0.01% and 0%). The results (obtained using the Mathematica computer program) are shown in Table 5.2.

As expected, the higher the number of populations, the lower the error of the estimated %K, and we can choose a CL closer and closer to 80%. Note also that each error bar is within the previous one, as it should (because an estimation with more populations, as designed above, is obviously more precise). From the last column we can safely conclude, with 80% CL, that the percentage of haplogroup K in a population for which we have measured a sample of 10 "0" individuals (i.e., in which none of the 10 individuals has the haplogroup K), is within the range 0%-14%. Thus we have applied the error bar 0-14%K to the sample 'Portugal coastal Early Neolithic' in the main paper.

Number of possible populations	upper limit of the % of haplogroup K	error bar of the % of haplogroup K	confidence level (CL)
11	(10±10)%K	0%-20%K	90.40%
101	(13±1)%K	0%-14%K	80.80%
1,001	(13.6±0.1)%K	0%-13.7%K	80.21%
10,001	(13.61±0.01)%K	0%-13.62%K	80.02%

Table 5.2 Error bar estimation for the regional culture 'Portugal coastal Early Neolithic' (10 individuals and 0%K).

Our method could be also applied to cases in which the haplogroup percentage is different from 0%, but calculations would be more tedious (because it is less straightforward to compute, e.g., the probability of 4 "0"s and 6 "1"s than to compute that of 10 "0"s). If the haplogroup percentage is different from 0%, we prefer to use the bootstrap approach because it is a reasonable method which makes it possible to compare directly to the error bars estimated by other authors (e.g., references [78, 176]).

Besides the sample 'Portugal coastal Early Neolithic' (which has 10 individuals, none of them with haplogroup K), in Fig. 5.2 (main paper) there is another sample with 0%K, namely 'Romanian Late-Middle Neolithic' (which has 9 individuals, none of them with haplogroup K). Repeating the same procedure as above for 9 (instead of 10) individuals, the result is that, with 80% CL, the percentage of haplogroup K in the original population is 0-15%K. The upper bound is higher than for 10 individuals, as it should, because with fewer individuals in a sample, inference about properties of the complete population (from which the sample has been drawn) is obviously more uncertain.

5.8.11. Text S11. A more complicated simulation model

In our main paper and in Sec. 5.8.5, the equations used to compute cultural transmission assume that both male and female hunter-gatherers are equally liable to form mixed couples with Neolithic individuals. However, ethnographic studies show that, in similar situations, mating takes place mostly between female hunter-gatherers and male farmers (see, e.g., reference [208]). If only female hunter-gatherers can mate with farmers, then none of the HN couples will contribute haplogroup K to the Neolithic gene pool, because mtDNA is inherited only from the mother. Note, however, that taking this point into account will not modify the genetic contribution of HX couples (because none of the parents has haplogroup K) neither NX couples (because both N and X are farmers, so the female can be either of them), neither of course HH, NN nor XX couples. On the other hand, the maximum possible number of both HN and HX couples will be smaller (by 50%) than in the model in the main paper and Sec. 5.8.5, because only female HGs can take part in them. Therefore, some genetic impact could in principle be expected if using this more realistic approach. Here we take this point into account, by means of an alternative cultural transmission scheme described below. We find, however, that the change in the results is in fact minimal, so the conclusions in the main paper do not change. We will also suggest some intuitive explanations of why this effect is so small.

In this model (Program S2, available at the journal web or at http://copernic.udg.es/QuimFort/2017_08_07r__Program_S2.zip) only part of the hunter-gatherer population (the females) can mate with farmers, so we have to consider separate sub-populations for men and women. Let $M_{HG}(x,y,t)$ and $W_{HG}(x,y,t)$ stand for the number of hunter-gatherer men and women, respectively, present in a cell after dispersal (step 1 in the main paper, Sec. 5.4). Therefore, the total number of hunter-gatherers in the cell is $P_{HG}(x,y,t) = M_{HG}(x,y,t) + W_{HG}(x,y,t)$. Likewise, let $M_F(x,y,t)$ and $W_F(x,y,t)$ stand for the farmer sub-populations of men and women, which are in turn divided into $M_N(x,y,t)$ and $W_N(x,y,t)$ for the farmer population with haplogroup K present in the cell, and $M_X(x,y,t)$ and $W_X(x,y,t)$ farmers that do not have haplogroup K. In the computer code we assume that initially there is gender balance in all populations, i.e. that there is the same number of males and females, and that in the new generations there is also equal probability to be born male or female.

5.8.11.1. Cultural transmission

The cultural transmission process (step 2 in the main paper, Sec. 5.4, and detailed in Sec. 5.8.5) is now replaced by the following.

Cross-matings between cultural groups. For cells with Mesolithic and Neolithic individuals, we first compute the mixed couples by taking into account that only hunter-gatherer women can mate into the farmer community. Let us first find, for example, the probability for a hunger-gatherer woman to mate with a farmer man who has haplogroup K. Under random mating (same tendency to mate with a hunter-gatherer man than with a farmer man), this probability would be simply the fraction of men with haplogroup K (relative to the whole male population in the cell). However, in general, this probability will be reduced by the interbreeding parameter η which, when $\eta < 1$, favors mating within the same population over mixed matings. Therefore, the probability for a hunger-gatherer woman to mate with a farmer man who has haplogroup K is given by [94]

$$\eta \frac{M_N}{M_N + M_X + M_{HG}}, \tag{S43}$$

where $M_N + M_X + M_{HG}$ is the total male population in the cell. Multiplying this probability (S43) by the number of hunter-women, W_{HG}, we find the corresponding number of mixed couples

$$couples\ M_N W_{HG} = \eta \frac{W_{HG} \cdot M_N}{M_N + M_X + M_{HG}}. \tag{S44}$$

Similarly, we find for the number of matings to farmer men who do not have the haplogroup K

$$couples\ M_X W_{HG} = \eta \frac{W_{HG} \cdot M_X}{M_N + M_X + M_{HG}}. \tag{S45}$$

Note that equations (S44)-(S45) are similar to equations (1)-(2) in the main paper, so we are actually applying vertical cultural transmission, but only to the subgroups liable to form mixed couples.

Analogously to the model used in our main paper (equation (S7)-(S9)), we next compute the number of farmer men and hunter-gatherer women who do not take part in the mixed matings above,

$$M'_N = M_N - couples\ M_N W_{HG} \tag{S46}$$

$$M'_X = M_X - couples\ M_X W_{HG} \tag{S47}$$

$$W'_{HG} = W_{HG} - couples\ M_N W_{HG} - couples\ M_X W_{HG} \tag{S48}$$

Cross-matings between genetic groups of farmers. We now compute the number of couples between farmer individuals of different genetic groups (N and X). Since some farmer men have mated with hunter-gatherer women, we now have fewer farmer men than farmer women (remember that we initially had gender balance). We can find the probability for a farmer man to mate with a farmer woman of the other genetic group. This will now just be the fraction of women of the other genetic group (relative to all farmer women). As argued above equation (3) in the main paper, there is no reason to assume any preference toward or against matings within the same genetic group, and therefore we can assume $\eta = 1$ (random mating). As a result, the number mixed genetic couples within the farmer community are given by

$$couples\ M_N W_X = \frac{M'_N \cdot W_X}{W_N + W_X}, \tag{S49}$$

$$couples\ M_X W_N = \frac{M'_X \cdot W_N}{W_N + W_X}, \tag{S50}$$

where $W_N + W_X = W_F$ is the total number of farmer women. Equations (S49)-(S50) are analogous to equation (S10) for the simpler model used in our main paper.

***Matings within groups*.** Finally, the number of couples between farmers of the same genetic group is constrained by the number of unmated men (which are fewer in number than unmated women). In the same way, the number of couples between hunter-gatherers is constrained by the number of unmated women. Therefore,

$$couples\ M_N W_N = M'_N - couples\ M_N W_X\ , \tag{S51}$$

$$couples\ M_X W_X = M'_X - couples\ M_X W_N\ , \tag{S52}$$

$$couples\ M_{HG} W_{HG} = W'_{HG}. \tag{S53}$$

Note that, in contrast to the analogous equations in the simpler model applied in the main paper [equations (S11)-(S13)], here we do not need to divide by two because we are now dealing with men and women separately.

5.8.11.2. Reproduction

The following scheme replaces the reproduction step in the main paper (Sec. 5.4) and Sec. 5.8.5. We apply the following rules. (i) Each couple will have $2R_{0,i}$ children, because $R_{0,i}$ is computed per individual and there are two individuals per mating. However, the net growth rate $R_{0,i}$ is different for farmers and HGs ($i = F, HG$). Applying that the children from cross matings between HG and F will be farmers [52, 208], we use $R_{0,HG} = 1$ for matings in which both parents are HGs (assuming that the HG population is stationary), and $R_{0,F} = 2.45$ [183] for HN, HX, NN, XX and NX matings. (ii) Since mtDNA is inherited from the mother, all the children from each couple will become part of the same genetic group as the mother. (iii) We assume equal probability for the children being male or female, so 50% of the new population will be men and the other 50% women. Under these three rules, the number of men and women in the next generation is given by

$$M_{HG}(t+1) = W_{HG}(t+1) = R_{0,HG} \cdot couples M_{HG} W_{HG}\ , \tag{S54}$$

$$M_N(t+1) = W_N(t+1) = R_{0,F}(couples\ M_N W_N + couples\ M_X W_N)\ , \tag{S55}$$

$$M_X(t+1) = W_X(t+1) = R_{0,F}(couples\ M_X W_X + couples\ M_N W_X + couples M_X W_{HG} + couples M_N W_{HG})\ . \tag{S56}$$

These equations are analogous to equations (S14)-(S16) for the simpler model applied in the main paper. Note that the couples *HN* appear in equation (S15) but not in equation (S55), because here all HGs in those matings are women and their mtDNA haplogroup is inherited by their children (so none of the latter will have haplogroup K and, therefore, never belong to population N but always to X). Finally, although this is not necessary to perform the simulations, using equation (S46)-(S48) into (S51)-(S53) and the results into (S54)-(S56) we can relate the population numbers at generation $t+1$ to those at the previous generation t,

$$M_{HG}(t+1) = W_{HG}(t+1) = R_{0,HG}\ (W_{HG}(t) - couples M_N W_{HG} - couples M_X W_{HG})\ , \tag{S57}$$

$$M_N(t+1) = W_N(t+1) = R_{0,F}\ (M_N(t) - couples\ M_N W_{HG} - couples\ M_N W_X + couples\ M_X W_N\ ), \tag{S58}$$

$$M_X(t+1) = W_X(t+1) = R_{0,F}\,(M_X(t) - couples\ M_X W_N \qquad \text{(S59)}$$
$$+couples\ M_N W_X + couples\ M_N M_{HG})\,,$$

which are analogous to equations (S17)-(S19) for the simpler model used in the main paper.

5.8.11.3. Simulation results

If we apply this new cultural transmission-reproduction scheme to the same initial conditions as in the main paper, we obtain basically the same results, as can be observed by comparing Fig. 5.22 below to Fig. 5.15 (i.e., Fig. 5.3 in the main paper). The absolute differences between the predicted fractions of individuals with haplogroup K are lower than 0.002. Therefore, although the scheme used in the main paper is more simplified, its results are close enough to those obtained here to validate the use of such an approximation. Also, because both models yield nearly the same results, the conclusions of the paper remain unchanged.

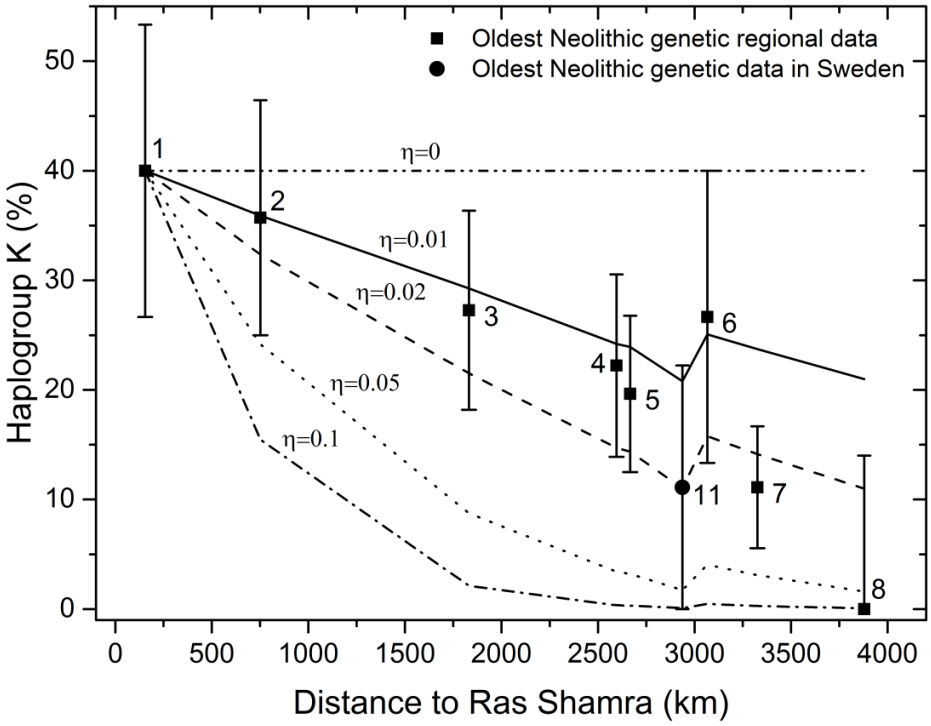


Figure 5.22 This figure is the same as Fig. 5.15 (i.e., Fig. 5.3 in the main paper), but applying the more refined model in Sec. 5.8.11. It shows the percentage of mtDNA haplogroup K present in the farmer population as a function of distance to Syria. Black squares correspond to the measured data. Lines correspond to the results of the simulations, using the model in Sec. 5.8.11, for several values of the interbreeding parameter η. The results that follow from this more precise model are almost the same as those from the model used in the main paper (compare this figure to Fig. 5.15, i.e. Fig. 5.3 in the main paper).

The fact that the refined model (this section) and the approximate one (main paper and Sec. 5.8.5) lead to much the same results does not seem very surprising, for the following reasons. It is true that in the more refined model (this section) only female hunter-gatherers (without haplogroup K) are incorporated into the farming populations (thus all of their children lack haplogroup K), whereas in

the approximate model (main paper and Sec. 5.8.5) additional (male) hunter-gatherers are also incorporated. However, the children of the latter do not always have haplogroup K (they will have it if the mother belongs to group N, but not if she belongs to group X). Thus the *HN* matings that are not taken into account in the more refined model (this section) lead not only to children who have haplogroup K, but also to children who do not have in the approximate model (main paper and Sec. 5.8.5). Then it seems reasonable that the effect of this refinement on the percentage of haplogroup K is small. Moreover, the genetic contribution of most matings (i.e., HX, NX, HGHG, NN and XX) is unaffected by this refinement in the model. Another difference is that the maximum possible number of HN and HX matings is lower in the model in this section (because only women HGs and men farmers take part in them), but again all other matings (i.e., NX, HGHG, NN and XX) are unaffected and, moreover, in our simulations we have observed that the percentage of haplogroup K becomes almost constant many generations before all possible matings with HGs have taken place (i.e., before the local HGs extinguish in the model in the main paper and Sec. 5.8.5, or before the local HG women extinguish in the model in this section).

5.8.12. Text S12. Approximate, one-dimensional model

In this work we have concluded that a value of the interbreeding parameter η as low as $\eta = 0.02$ (which is very small, as compared to the maximum possible value $\eta = 1$ [94]) explains the observed cline of haplogroup K in aDNA data (main paper, Fig. 5.3). In order to perform a check to the validity of this new result, we conceived an approximate, simpler model as follows. The model in the main paper (detailed in Sec. 5.8.5), as well as the more elaborate one in Sec. 5.8.11 above, considers a two-dimensional (2D) grid, and distinguishes sea, mountain, coast and inland cells to simulate a real map of Europe. On this 2D grid, individuals are exchanged between cells via *sea* travels (up to 150 km, as implied by archaeological data; see Sec. 5.8.6) and also via *inland* travels (of 50 km, as implied by ethnographic data [139]). We reasoned that, since *sea* travels can be substantially longer than *inland* travels, a one-dimensional (1D) model (representing the Mediterranean coast) could be a simpler way to describe roughly the dynamics of the system. Although this is admittedly a simplification, and will obviously lead to less precise results, it seems reasonable to expect that it can be useful to check the main conclusion of our work (namely, that $\eta \ll 1$, as explained above).

In this one-dimensional model (Program S3, available at the journal web or at http://copernic.udg.es/QuimFort/2017_08_07r__Program_S3.zip) we assume a line of 150 nodes, each one separated 150 km from their two neighbors. This corresponds to a total of 22,350 km between the two extreme nodes (150 km multiplied by 149 jumps between nodes). As in the main model, initially only one node has Neolithic population (3,200 individuals) but no Mesolithic population, and all other nodes have no Neolithic population and 160 Mesolithic individuals per node (values obtained from ethnographic data[13] and the area of a cell in the main model). From the node with Neolithic individuals, the Neolithic population expands along the line (which corresponds to the Mediterranean coast) by performing, each generation, the steps of population dispersal, cultural transmission and reproduction. The latter two steps are treated here in the same way as in the main paper (detailed in Sec. 5.8.5). Dispersal, on the other hand, needs to be treated differently because of the unidimensionality of space.

5.8.12.1. Dispersal

The nodes in the 1D grid (this section) are equivalent to coastal nodes in the 2D grid (main paper and Sec. 5.8.5). The population present at a coastal cell in the 2D model can stay with a 38% probability (persistence) [139], or it can travel either inland or by sea (with the number of individuals taking each route depending on the number of sea neighbors). In general, in the 2D model we have three possibilities (we ignore the cases where a neighbor is a mountain cell), depending on the number of sea neighbors: (i) one sea neighbor implies that 25% of the population that travels (15.5% of the total population, computed as $(1/4)(1-p_e)$) moves by sea, (ii) two sea neighbors means that 50% of the traveling population (31% of the total population, computed as $(1/2)(1-p_e)$) will travel by sea, and (iii) three sea neighbors means that 75% of the traveling population (46.5% of the total population, computed as $(3/4)(1-p_e)$) will travel by sea.

In the approximate 1D model (this section), we are only considering sea travel. In contrast, in the 2D model (main paper and Sec. 5.8.5) we consider both sea and inland travel. For this reason, obviously in the 1D model if we allowed for all of the population that can travel (62% of the cell population) to migrate by sea, the speed of the front would largely overestimate the results obtained with the more realistic 2D model (which agree with the archaeological data, see Sec. 5.8.6). In addition, nodes in the 1D model are separated 150 km, so shorter jumps are not possible. In contrast, in the 2D model not all of the population that travels by sea moves 150 km away from the origin, because part of it moves to closer coastal locations. Therefore, for the 1D model to provide equivalent results, it is important that a lower fraction of the population travels by sea, so that the results are realistic. This implies that part of the population has to disappear from the system in the 1D model, in order to take care of the fraction of the population that travels inland (and, therefore, does not contribute to the coastal expansion) in the 2D model. Hence, in our 1D model a fraction $\alpha(1-p_e)$ of the population travels by sea, and a fraction $(1-\alpha)(1-p_e)$ of the population disappears. We find by trial and error a fair approximation to the value of α by setting the following constraint. For the 1D model to be a good approximation of what happens in a real geography, the arrival times for the 1D and the 2D models must be the same. We have chosen a coastal cell as a test origin for the 2D model, and a cell located 5,100 km away (distance measured along the coast) to calibrate the 1D model (5,100 km corresponds to 34 jumps of 150 km each). As in Sec. 5.8.6, the arrival time of the Neolithic to a cell is recorded by the simulations as the time when the population of farmers is 10% of its saturation value. In the 2D model, and with jumps of 150 km, a node located at 5,100 km is reached within 75 generations. With the 1D model, a node located 5,100 km away from the origin (i.e., 34 cells away) is reached within 52 generations if we assume that $(1/2)(1-p_e)$ individuals travel by sea, within 69 generations if $(1/3)(1-p_e)$ individuals travel by sea, and within 83 generations if we assume that $(1/4)(1-p_e)$ individuals travel by sea. This allows us to fine tune the best approach to a fraction of the population that must travel by sea in the 1D model to $0.3(1-p_e)$, which yields an arrival time within 75 generations, equivalent to the one measured in the 2D model.

Therefore, in the dispersion process of the 1D model, 38% of the population stays in the same node, and a fraction 0.3 of the remaining population ($0.3(1-p_e)$, i.e 18.6% of the total population) will travel by sea, half of them forward and the other half backward (similarly to the 2D model, where all possible destinations receive equal fraction of the sea travelling population). The rest of the population, as mentioned above, disappears from the system, representing the population that would travel inland (in the 2D model).

5.8.12.2. Simulation results

We now run the 1D model (Program S3) and the 2D one (Program S1), under the initial condition observed from ancient mtDNA data in Syria, namely that 40% of the initial farmer population has haplogroup K, and we compare the results at several distances from the origin. For the 2D model, we choose a coastal node as origin and measure the distances along the coast, rather than with straight lines.

We show the results for several values of η in Fig. 5.23, where we have measured the fraction of the population with haplogroup K at several locations, 10 generations after the local Neolithic arrival (according to the simulations). From Fig. 5.23 we can see that the 2D model (lines + symbols) always predicts a lower fraction of population with haplogroup K than the 1D model (lines). However, given that the 1D model is just an approximation, it is interesting to see that the results from both models have similar behaviors and are close enough, so the 1D model is a useful check of the results of the 2D model (especially, the conclusion that very low values of η are necessary in order for the genetic cline to extend across a distance similar to that from Syria to Portugal).

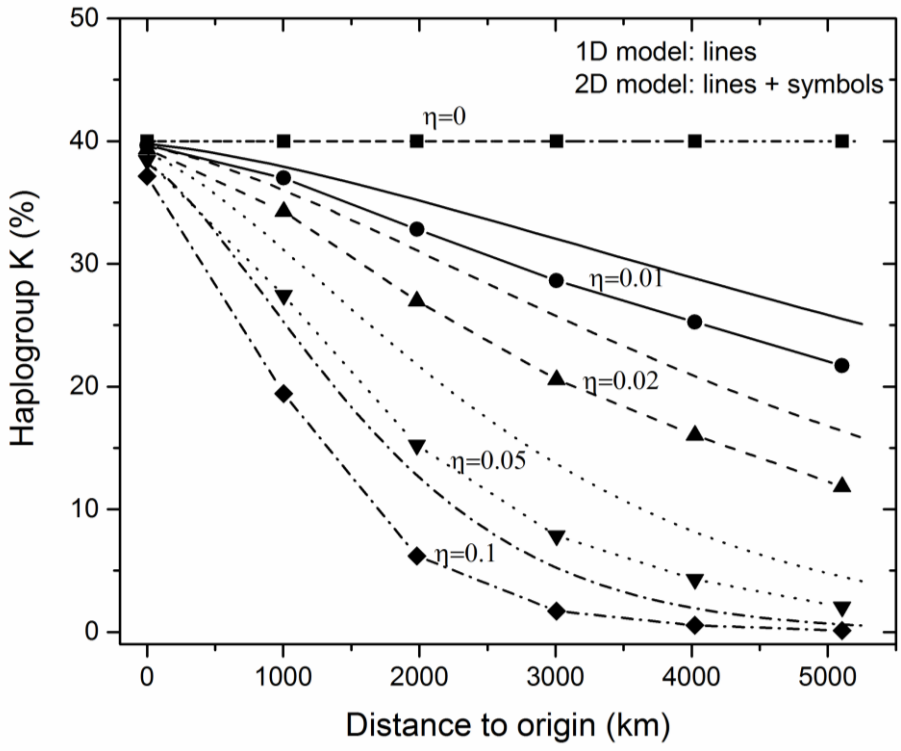


Figure 5.23 Percentage of mtDNA haplogroup K present in the farmer population that disperses along the coast, as a function of distance to an origin coastal node. Lines correspond to the approximate 1D model developed in Sec. 5.8.12. Lines with symbols correspond to the 2D model on a real map of Europe used in the main paper. All results are measured 10 generations after the local arrival of the Neolithic front (according to the simulations).

5.8.13. Text S13. The speed of waves of advance in homogeneous space

In order to perform a check of our simulations we recall that, in two-dimensional homogeneous space (i.e., without seas neither mountains), the speed of the waves of advance of farmers corresponding to our reproduction-dispersal-interbreeding scheme is [94]

$$speed = min_{\lambda>0} \frac{ln\{R_{0,F}(1+\eta)[p_e + (1-p_e)I_0(\lambda r)]\}}{T\lambda}, \quad \text{(S60)}$$

where $I_0(\lambda r)$ is the modified Bessel function of the first kind and order zero, r is the average distance that an individual moves per generation, $R_{0,F}$ is the net reproduction rate for farmers, p_e is the persistence, and T is the generation time. Figure 5.24 shows the results (lines) obtained from equation (S60) when using the same values as in the main paper, Sec. 5.4, i.e. $r = 50$ km [52, 139], $R_{0,F} = 2.45$ [183], $p_e = 0.38$ [139] and $T = 32$ yr [182]).

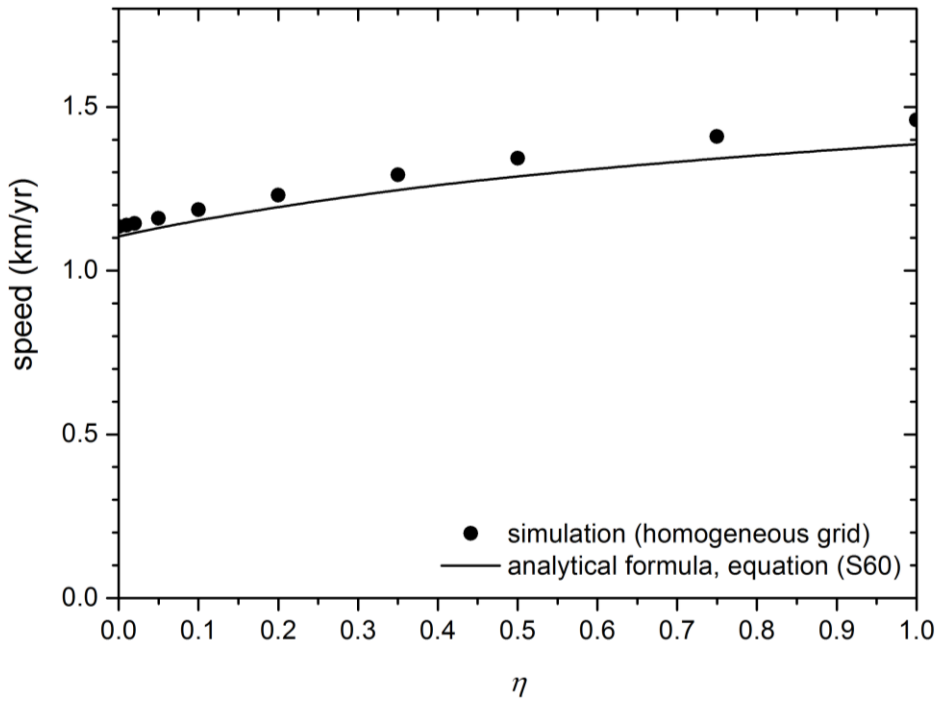


Figure 5.24 Predicted front speed from the computational model (Program S5) and an analytical approximation on a homogeneous grid. Our simulations on a homogenous grid, i.e. without seas nor mountains (symbols), agree with and the corresponding analytical formula, Eq. (S60) (curve). This is a useful check of our simulations. All results have been obtained using $r = 50$ km [52, 139], $R_{0,F} = 2.45$ [183], $p_e = 0.38$ [139] and $T = 32$ yr [182].

We performed additional simulations using, instead of a map of Europe, a homogeneous grid of land nodes (i.e., without sea-travel neither mountain barrier effects), as we did in Fig. 5.21. However, in order to compare to equation (S60), now we will analyze the spread rate of the front (not the genetic cline as in Fig. 5.21). We perform our simulations with Program S5 (available at the journal web or at http://copernic.udg.es/QuimFort/2017_08_07r__Program_S5.zip), which performs the same logic as Program S1 (Sec. 5.8.5), but on a homogeneous grid where the Neolithic spreads from its center

(Program S5 differs from Program S4, Sec. 5.8.8, in how the initial conditions are set). Initially there are hunter-gatherers (at their saturation density) in all cells but the central one, where there are only farmers (also at their saturation density). Since we are now only interested in the arrival time, the genetic composition of the initial farming population does not affect the results, so we set it at 100%K. In each simulation, a wave of advance of farmers propagates outwards from the center of the grid. In order to determine its speed, for each cell along the x-axis we record the time when the farmer population reaches a population number equal to 10% of its saturation value (however, the wave-of-advance speed is not affected by this percentage, i.e. we would obtain the same speeds by using, e.g., 90%). The speed is then computed as the slope of the linear fit of the distances from the origin versus arrival times. Figure 5.24 shows the speed of the waves of advance along the horizontal direction (symbols), obtained from those simulations, as a function of the interbreeding parameter η (with $0 \leq \eta \leq 1$, see main paper, Sec. 5.4, Cultural transmission). Errors are below 6% (see Data S5), which is reasonable because, in contrast to equation (S60), which assumes a continuous space, simulations are necessarily performed on a grid, i.e., using only a finite number of spatial locations. This confirms the validity of our simulations.

5.8.14. Text S14. Pre-Neolithic haplogroups in Neolithic communities

The analysis performed in the main paper indicates, based on the variation of the mitochondrial haplogroup K, that the Neolithic expansion was mostly demic, although with a low contribution of cultural diffusion. Under these circumstances, in addition to the decay in the presence of haplogroup K, we should also be able to observe an increase in the presence of hunter-gatherer haplogroups in the Neolithic communities. Mitochondrial DNA from hunter-gatherers in Central-European was limited to haplogroups U, U4, U5 and U8 [78, 203, 204, 252]. These lineages showed also a high frequency among the western Mediterranean hunter-gatherers [167, 253], but the latter also presented important frequencies of haplogroup H lineages [78, 167]; especially haplogroups H1 and H3, which are related to a post glacial expansion from an Iberian refugium [166, 253, 254] (in central Europe, on the other hand, H lineages are linked to the spread of the Neolithic [255]).

Figure 5.25 shows the percentage of hunter-gatherer haplogroups (i.e., haplogroups U, U4, U5 and U8 for regions in the Near East and Eastern and Central Europe, and haplogroups U, U4, U5, U8, H1 and H3, for regions in Iberia and Southern France) with their error bars, for the same regions as in Fig. 5.2 in the main paper. In Fig. 5.2, we have fitted a straight line because this is the simplest fit such that it crosses all error bars of the oldest Neolithic cultures (squares and circle). However, this is not possible for Fig. 5.25, so we fit more appropriate curves, namely a power and an exponential function. This is reasonable since, as we mentioned in the main text, there is no reason why a genetic cline should be linear, and we could actually also fit a decreasing power or exponential curve to Fig. 5.2. We see that both fits in Fig. 5.25 show that the percentage of hunter-gatherer lineages present in the Early Neolithic populations increases with distance once the Neolithic front reaches the Central European area (Region '3 Hungary-Croatia Stracevo'), which agrees with our hypothesis that the hunter-gatherer contribution to the Neolithic pool would have increased away from the origin of expansion.

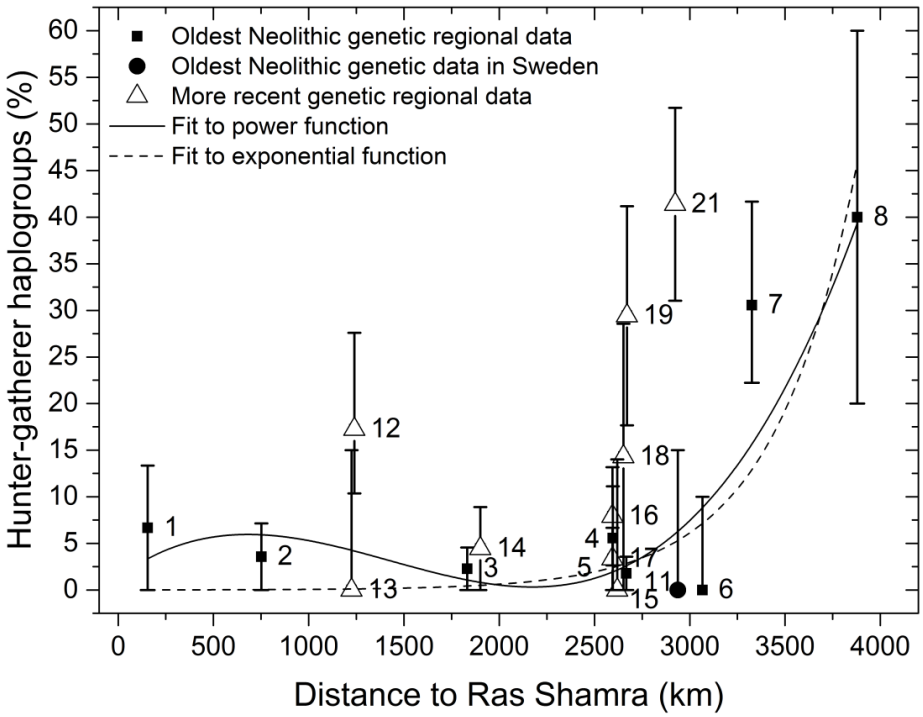


Figure 5.25 Observed percentage of hunter-gatherer mtDNA haplogroups as a function of the great-circle distance from Ras Shamra (Syria). The haplogroups considered in all regions are U, U4, U8 and U8, while haplogroups H1 and H3 are also included in western Mediterranean regions: 6 North-Eastern Spain Cardial, 7 Spain Navarre, 8 Portugal coastal Early Neolithic and 21 South-Eastern France Treilles. Each number denotes the same cultures as in Fig. 5.1 (as in Fig. 5.2, regions with fewer than 8 individuals have been ignored to avoid very large error bars). The solid and dashed lines are regression fits to the 8 oldest regional data (squares) and the oldest data in Sweden (circle). Error bars display 80% confidence-level intervals (see Sec. 5.4.2).

In Fig. 5.25, for distances below 2,000 km the two considered fits show different behaviors, both consistent with the data and their error bars, and both yielding a similar goodness of fit. Therefore, it is not possible to establish which fit is more reasonable. But this does not change our conclusion that the percentage of HG haplogroups increases with distance, as expected. Our results in the main paper attempt to provide an estimation of the average intensity of cultural diffusion at the continental scale, i.e., our purpose is to analyze the overall process, not regional differences. However, it is worth to note that recent studies have suggested that the effect of cultural diffusion increased as farmers spread to further locations [207], which would agree nicely with our observations in Fig. 5.25.

From Fig. 5.25 we can also see that, in general, in later periods (triangles) the presence of hunter-gatherer haplogroups increases, since most triangles are located above the lines fitting the data for Early Neolithic populations (black squares and circle). This behavior is consistent with the conclusion in our main paper that after the first arrival, the farmer populations continue incorporating local hunter-gatherer individuals, and therefore the presence of hunter-gatherer haplogroups in the Neolithic populations should increase (as observed in Fig. 5.25).

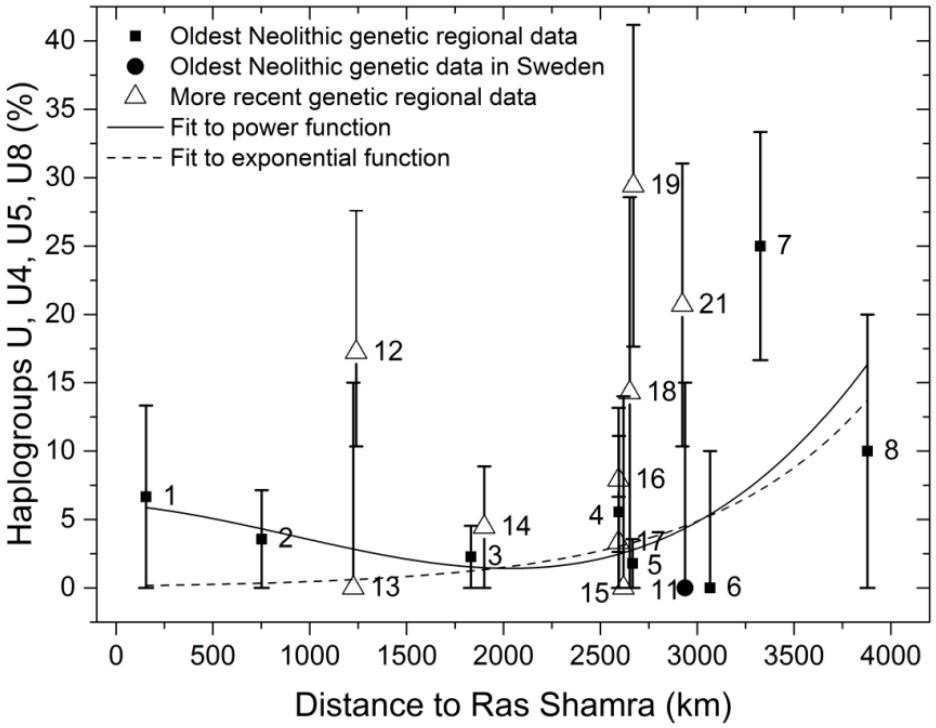


Figure 5.26 Observed percentage of U haplogroups in Neolithic populations as a function of the great-circle distance from Ras Shamra (Syria). Labels denote the same cultures as in Fig. 5.1 (as in Fig. 5.2, regions with fewer than 8 individuals have been ignored to avoid very large error bars). The solid and dashed lines are regression fits to the 8 oldest regional data (squares) and the oldest data in Sweden (circle). Error bars display 80% confidence-level intervals (see 5.4.2).

We would like to stress that the observed increase at longer distances is not an artificial effect of including H lineages; if we considered only U lineages we would again obtain an increase of hunter-gatherer haplogroups in the early farmer populations, as shown in Fig. 5.26.

PART III

Discussion and conclusions

6. Discussion

The main work of this thesis encompasses three different systems that can be described using similar mathematical models, which deal with population diffusion (or dispersal) and reaction (reproduction and interactions). These systems have been presented as a collection of papers (Chapters 3-5). Each one includes an introduction to the problem, the mathematical models we have used, computational analyses, results and conclusions. Finally we review the main results that we have found along these three Chapters, and we discuss their similarities.

6.1. Mathematics behind viral replication and applications

The mathematical analyses in Chapters 3 and 4 are based on very similar models of viral infections. In these two Chapters, our main purposes have been to calculate front propagation speeds and to compare them to experimental data.

In the first system (Chapter 3), T7 viruses infect *E. coli* bacteria *in vitro*. This system was discussed already in 1945 by M. Demerec and U. Fano [256], and has attained some interest in Physics because it has allowed to test time-delayed front propagation theories in the laboratory [4].

The second system (Chapter 4) aims to go one step further by applying these ideas to a system with a real clinical application, namely the treatment of GBM tumors with VSVs. Although the model is simple (radial symmetry, no vasculature, infection starting from a single point, etc.), Chapter 4 provides a quantitative framework to describe the spread of GBMs and VSVs infecting them.

For the system analyzed in Chapter 3, Yin and McCaskill [14] had previously developed a model which we have already detailed in Sec. 1.2.3, Eqs. (1.19)-(1.21). In fact, this model can be recovered from our main model, which is described by Eqs. (3.4)-(3.6), if assuming a vanishing time delay, i.e. $\tau = 0$. In their original work, Yin and McCaskill [14] noted that, if using realistic parameter values, the observed speeds were substantially faster than those predicted by their model. They could obtain agreement between predicted and observed speeds only by fitting three adjustable parameters in their model (Fig. 3b in Ref. [14]). In contrast, in Chapter 3 we have not fitted any parameter but used realistic values for all parameters, obtained from independent experiments. We have seen, again, that using the Yin-McCaskill model (Eqs. (1.19)-(1.21)) the speeds calculated for all 3 strains of the T7 virus are much faster than the experimental ones (Fig. 3.2). This is because the Yin-McCaskill model does not take into account that, after the virus infects a bacterium, it takes a certain time τ to complete the lytic cycle (see Fig. 1.2). During this time interval τ, the virus cannot move because it is inside a cell (which is immobilized by agar). In the Yin-McCaskill model, this delay time is not taken into account, and this is why the resulting velocities are faster than the experimental ones (Fig. 3.2). In contrast, in our new model [Eqs. (3.4)-(3.7)] we take into account this delay time both in the diffusion [terms proportional to τ in Eq. (3.5)] and in the reaction [last term in Eqs. (3.4), (3.5) and (3.7)]. In this way, our predicted speed is substantially closer to the observed one for all 3 strains (Fig. 3.2).

Previous models had already included the delay time effect into either the diffusive of the reactive terms, but not into both terms. On one hand, previous models that took into account the delay time in the diffusion [4, 21, 145] did not do so in the reaction and used instead a logistic term (see Eq. (3.2)) according to which the death rate of infected cells would be proportional to the free space, which is not biologically reasonable. On the other hand, recent models by Gourley and Kuang [146] and Jones et al. [147] include the time delay only in the reaction (death of infected cells and creation of new viruses), but not in the diffusion. Moreover, both studies [146, 147] assume that all cells infected at time $t-\tau$ die exactly at time t. This does not agree with one-step experiments because in those experiments, not all viruses appear at the same time after infection. Indeed, some viruses appear 14 min and others 21 min after infection (Fig. 3.1). The main difference between those two previous works is that Jones et al. [147] assumed that *all* infected cells release new viruses, while Gourley and Kuang [146] include an additional parameter $e^{-\mu_I \tau}$ which accounts for the infected cells that may have died without contributing to virus replication. This is why the Gourley-Kuang model yields slightly slower front propagation speeds than the model due to Jones et al. (Fig. 3.2). In any case, both models [146, 147] contradict one-step experiments. In contrast, in our model infected cells present at time $t-\tau$ begin to die at time t, and do so gradually thereafter, which agrees with one-step experiments (Fig. 3.1). With our new approach, and taking also into account the effect of the delay time on the diffusive process, we achieve better agreement between predicted and front speeds than the Yin-McCaskill [14], Gourley-Kuang [146] and Jones et al. [147] models (Fig. 3.2).

In Table 6.1 we illustrate the models reviewed above by comparing their predicted speeds for the T7 wild strain infecting *E. coli* to the observed speed (namely, $c_{obs} = 0.195 \pm 0.012$ mm/h [15]). The lowest error (9%) is attained for our model. This indicates that a satisfactory description of this system, both at the quantitative and at the conceptual level, requires including both reactive and diffusive time-delay effects.

	experimental [15]	Yin and McCaskill [14]	Jones et al. [147]	Gourley and Kuang [146]	de Rioja et al. (Chapter 3)
speed [mm/h]	$0.195^{+0.012}_{-0.012}$	$0.375^{+0.054}_{-0.078}$	$0.261^{+0.020}_{-0.035}$	$0.254^{+0.020}_{-0.035}$	$0.212^{+0.005}_{-0.011}$
Relative errors [%]	-	92%	34%	30%	9%

Table 6.1 Front propagation speed and error (relative to the experimental speed) for the T7 wild strain infecting *E. coli*, according to four reaction-diffusion models. The model by Yin and McCaskill ignores the delay-time effect, those by Jones et al. and by Gourley and Kuang take it into account but only in the reactive process, and the model in Chapter 3 takes it into account both in the reactive and in the diffusive processes.

Chapter 4 applies models very similar to the ones developed in Chapter 3 to obtain front propagation speeds of viruses and tumors. The main goal of this study is to give mathematical support to this type of oncolytic treatment. Three increasingly complex mathematical models have been presented, and we have compared their results with observed data. Model 1 is based on a previous one by Wodarz et al. [45], but we do not assume that the virus concentration is stationary, i.e. we use three evolution equations (whereas Wodarz et al. [45] used two equations, see Sec. 1.2.4). This model 1 is a non-delayed model (such as the one by Yin & McCaskill reviewed above [14]), so the resulting speeds will

necessarily be faster than those from models that take into account the delay time. Model 2 is time-delayed but only concerning reaction, in the sense that it takes into account the fact that tumor cells do not die instantly, but rather take a time τ between being infected and killed by the virus. Finally, our model 3 takes into account the reactive time-delay effect (as model 2) and also the diffusive one, in the sense that during time τ viruses are inside the cells so they cannot freely move. The value of τ is not accurately known for the VSV-GBM system, but the *in vitro* experiments by Wollmann et al. [24] (Sec. 2.1.8) imply the range $2 < \tau < 12$ h. The experimental data (Sec. 2.1.9) imply the range of observed VSV speeds $c_{obs} = 4 - 5.4$ cm/h. Model 1 predicts an average speed that is two orders of magnitude faster than the observed range, namely $c_{model\,1} = 265$ cm/h (Fig. 4.2). So model 1 is inconsistent with the experimental data. For model 2, the predicted average speed is much slower, $c_{model\,2} = 10 - 22$ cm/h (Fig. 4.2). Still, for model 2 to agree with the observed range, we would need values of τ, Y or k_1 far from the real ones for the VSV-GBM system. So model 2 does not agree with the experimental data either. Finally, model 3 predicts speeds consistent with the experimental range, for τ between 5.0 h and 9.3 h (Fig. 4.2), namely $c_{model\,3}(5.0 \leq \tau \leq 9.3\text{ h}) = c_{obs} = 4 - 5.4$ cm/h. So again we see the importance of appropriately including the delay time τ in the diffusive and reactive terms of our models to reproduce the observed results.

Also in Chapter 4, the glioblastoma front speed has been calculated. This is much simpler mathematically than the three-species sets of equations that we have used to calculate virus infection speeds. Using the Fisher speed, Eq. (4.24), we have obtained $7.9 \cdot 10^{-5} < c_{GBM} < 4.33 \cdot 10^{-4}$ cm/h, which is in agreement with the *in vitro* speed range of $2.37 \cdot 10^{-4} < c_{GBM} < 5.54 \cdot 10^{-4}$ cm/h [153]. Note that the VSV speeds above are four orders of magnitude faster, which implies that the virus infection can eventually spread up to the border of the expanding glioblastoma and destroy it (see Fig. 4.1).

It is important to highlight that the delay time has a very important effect on virus spread rates. Therefore, in order to make trustable predictions, it is absolutely necessary to take it into account. Moreover, as we have seen in both Chapters 3 and 4, to obtain the best agreement between predictions and experimental observations, it is important that the predictions include the effects of the time delay both in the diffusive and the reproductive processes in virus infection systems.

6.2. Discovering the past through genetics and mathematics

We have collected all genetic data of Early and Middle Neolithic individuals that have been reported in the literature (Appendix A). In Chapter 5, we have used them to detect a clear cline of mitochondrial haplogroup K and we have proposed a mathematical model that can explain this cline.

In Fig. 5.1 we have plotted the mean dates and distances of Neolithic farmers belonging to the older regional cultures for which we have human genetic data. The space-time regression fit is highly linear $(R = 0.93)$. This means that the spread rate was approximately constant, as expected from wave-of-advance models (similarly to those used to describe the spread of viruses in Chapters 3-4).

The percentage of individuals with mitochondrial haplogroup K (%K) in our database decreases with increasing distance, but not linearly (error bars in Fig. 5.3). Nonetheless, this is not a problem because, in contrast to the space-time plot mentioned in the previous paragraph, there is no theoretical reason to expect a linear decrease for the percentage of a genetic marker versus distance. We have also noted

that the %K in each region tends to decrease over time (Fig. 5.2). Both decreases are expected intuitively, simply due to the incorporation of hunter-gatherers (who lack haplogroup K) into the populations of farmers (we have modelled this process using cultural transmission theory, see Eqs. (1.32)-(1.35)).

Analogously to the three-equation reaction-diffusion models with three species (viruses, uninfected cells and infected cells) that we have applied to viral spread (Chapters 3-4), in Chapter 5 we have used a model of three populations, namely hunter-gatherers (HG), farmers with haplogroup K, and farmers without haplogroup K. As explained in Sec. 2.3.3, most of the grid (representing the map of Europe and Near East) is first dominated by HGs (analogously to *E. coli* bacteria in Chapter 3 and to tumor cells in Chapter 4). Farmers (analogously to viruses in Chapters 3-4) start to spread from an initial point or region. At each location, farmers and HGs and farmers interact, in the sense that some HGs become farmers by cultural transmission (i.e. interbreeding or acculturation). This interaction leads to a decrease in the local number of HGs and an increase in the number of farmers, which is somehow analogous to the fact that the number of bacteria (Chapter 3) or tumor cells (Chapter 4) decrease when they are infected, and the number of viruses increases due to replication (Chapters 3-4). In this way, the Neolithic spreads (by land and sea) throughout the European continent.

The main goals of the simulations in Chapter 5 are to reproduce the observed cline (%K vs distance) and to estimate the relative importance of demic and cultural diffusion in the spread of the Neolithic across Europe. Figure 5.3 shows the result of these simulations for various values of the cultural transmission intensity η. For $\eta = 0$ (purely demic process), the %K remains constant in all regions of the map, which is at variance with the observed cline (error bars in Fig. 5.3). This indicates that both mechanisms (demic and cultural diffusion) played a role in the spread of the Neolithic. For $\eta \neq 0$, the percentage of haplogroup K tends to decrease with increasing distance from the Near East, as expected (because at larger distances farmers will have arrived at more regions saturated by HGs, so more HGs will have been incorporated). The higher the value of the cultural transmission intensity η, the more HGs are incorporated, and the steeper the cline (Fig. 5.3). Thus, by comparing the data (error bars) to the simulated clines for several values of η, we can estimate the value of η (Fig. 5.3). In this way, we have estimated that the intensity of cultural diffusion was $\eta \approx 0.02$. Therefore, cultural diffusion was remarkably weak, in the sense that only 2 out of every 100 farmers mated a hunter-gatherer or, alternatively, taught agriculture to a hunter-gatherer (Chapter 5). For this reason, we have argued that the most relevant process in the Neolithic spread in Europe was demic diffusion, i.e. the dispersal of populations, rather than cultural diffusion, i.e. the incorporation of HGs into the farming populations. One way to justify this quantitatively is to note that such a low value of η ($\eta \approx 0.02$) has very little effect on the spread rate [94]. This implies that cultural diffusion had a very small effect on the spread rate and, in this sense, we can say that the spread of the Neolithic across Europe was mainly demic.

7. Conclusions

This thesis has studied population range expansions of microbiological and human populations in spatio-temporal systems. By applying similar mathematical concepts to different systems, we have explained the front speeds of several virus strains, the dynamics of a virus-tumor system with medical applications, and an ancient genetic cline. We have also estimated the relative importance of demic and cultural diffusion in the Neolithic transition in Europe. Chapters 3, 4 and 5 use the delay time (or specifically, the generation time for human populations) in their respective models. Besides, in all three Chapters we have achieved a good agreement between theoretical models and observed data.

In Chapter 3 a new space-time reaction-diffusion model of virus infections has been presented, and it has been applied to study the front propagation speed of T7 virus through *E. coli* bacteria. This new model improves previous reaction-diffusion models, in which the evolution equation for infected cells lacked of biological meaning. The new mathematical perspective assumes that the death rate of infected cells is proportional only to its own density (evaluated a time τ before, which is the time between the infection and the first release of viruses), but not to the free space, providing a more understandable equation from a biological point of view. Thus, we have shown that it is important to incorporate the delay time into the reactive processes, not only in the diffusion process as in previous studies. Indeed, some previous models yield too fast speeds, whereas our equations successfully agree with *in vitro* results. We reach this conclusion not only by comparing to classical models (i.e. with no time delay), but also to recent models which include the delay time in the reactive but not in the diffusive process (Fig. 3.2). In Chapter 3 we have used T7-*E. coli* systems because there are quantitative experimental data for them, but our model should be useful to describe other virus infection systems.

The spread of a VSV (virus) infection in glioblastomas (tumor cells) is the problem that Chapter 4 aims to describe quantitatively by using reaction-diffusion equations. This subject (oncolytic virotherapy) has recently motivated many scientific fields of research because of its medical interest. Chapter 4 improves a previous model in three increasingly realistic steps, by generalizing and correcting some drawbacks. Again, we include time-delay effects in the reactive and diffusive terms. The last of our models includes all of the necessary items to satisfactorily explain the VSV-GBM system, namely three-population equations, infected tumoral cells that do not die instantaneously (reactive effect of the delay time), and viruses that do not move in space while they are inside infected cells (diffusive effect of the delay time). This last model turns out to be the only one yielding results in agreement with front speeds for oncolytic VSVs infecting glioblastomas, as observed *in vitro* (Fig. 4.2). Therefore, it is completely necessary to add the corrections of the delay time in the diffusive and reactive terms to fully describe real virus-tumor systems. Although Chapter 4 has focused on VSV-GBM system (because of the available experimental data), the mathematical model has been constructed by incorporating physically and biologically understandable equations and can be applied also to other virus-tumor systems.

In Chapter 5 we have analyzed the relative importance of demic and cultural diffusion in the Neolithic transition in Europe, from a reaction-diffusion computational perspective and based on archaeological and ancient genetic data. Using anthropologically realistic assumptions and parameter values, and a mathematical model that combines demic dispersal, population growth, and cultural transmission theory, we have built a computational model that explains how and why haplogroup K

displays a decrease in space and time, according to ancient genetic data. We have focused our attention on mtDNA haplogroup K because it is virtually absent in hunter-gatherer populations and has its maximum frequency in the Near East. We have performed different simulations, by varying the intensity of the interbreeding parameter η, i.e. giving more or less importance to the cultural diffusion relative to demic diffusion. Simulations make it possible to reproduce the observed cline, including a local minimum in Sweden (which can be easily explained because of the long sea travels, as compared to the shorter inland movements, and this also explains the faster spread along the Mediterranean that is observed from archaeological data). Finally, the simulation which best agrees with the observed genetic cline ($\eta \approx 0.02$) implies that 98% of farmers were not involved in cultural diffusion (Fig. 5.3). But it is important to highlight that the observed cline cannot be understood without cultural diffusion, since otherwise the percentage of haplogroup K would be uniform in space and constant in time. In conclusion, our research suggests that farmers involved in cultural diffusion, either by interbreeding (cross-mating) or by acculturation (teaching agriculture to hunter-gatherers), were a tiny but necessary fraction, namely about 2%. Similarly to Chapters 3-4, the ideas in Chapter 5 could be applied to other systems and lead to further insights in future research. Firstly, our approach could be applied to other genetic markers of importance in the study of the Neolithic spread across Europe [80], and perhaps different results for mtDNA (e.g., those in Chapter 5) and Y-chromosome markers could reveal different migratory behaviors of men and women (because mtDNA and the Y chromosome are maternally and paternally inherited, respectively). Secondly, our models could be applied to other human range expansions, if sufficient ancient genetic data become available (e.g., to the Neolithic spread in Asia [257]). And thirdly, it would be of interest to use integer numbers (i.e., population numbers instead of densities) in order to examine the possible role of drift effects in the formation of genetic clines [56].

PART IV

Bibliography and appendices

Bibliography

[1] T. Pujol and B. Comas, "Analytical expressions for the flame front speed in the downward combustion of thin solid fuels and comparison to experiments," *Phys. Rev. E,* vol. 84, no. 026306, 2011.

[2] S. J. Di Bartolo and A. T. Dorsey, "Velocity selection for propagating fronts in superconductors," *Phys. Rev. Lett.,* vol. 77, p. 4442, 1996.

[3] J. Fort, "Synthesis between demic and cultural diffusion in the Neolithic transition in Europe," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 109, pp. 18669-18673, 2012.

[4] J. Fort and V. Méndez, "Time-delayed spread of viruses in growing plaques," *Phys. Rev. Lett.,* vol. 89, p. 178101, 2002.

[5] S. Rendine, A. Piazza and L. L. Cavalli-Sforza, "Simulation and separation by principal components of multiple demic expansions in Europe," *Am. Nat.,* vol. 128, pp. 681-706, 1986.

[6] J. Fort and J. Pérez-Losada, "Can a linguistic serial founder effect originating in Africa explain the worldwide a phonemic cline?," *J. R. Soc. Interface,* vol. 13, pp. 1-9, 2016.

[7] V. L. de Rioja, J. Fort and N. Isern, "Front propagation speeds of T7 virus mutants," *J. Theor. Biol.,* vol. 385, pp. 112-118, 2015.

[8] V. L. de Rioja, N. Isern and J. Fort, "A mathematical approach to virus therapy of glioblastomas," *Biology Direct,* no. 11, 2016.

[9] N. Isern, J. Fort and V. de Rioja, "The ancient cline of haplogroup K implies that the Neolithic transition in Europe was mainly demic," *Sci. Rep.,* vol. 7, no. 11229, 2017.

[10] M. Delbrück, "Bacterial viruses or bacteriophages," *Biol. Rev. Camb. Philos. Soc.,* vol. 21, pp. 30-40, 1946.

[11] M. H. Adams, Bacteriophages, New York: Interscience, 1959.

[12] J. Yin, "A quantifiable phenotype of viral propagation," *Biochem. Biophys. Res. Commun.,* vol. 174, p. 1009–1014, 1991.

[13] A. L. Koch, "The growth of viral plaques during the enlargement phase," *J. Theor. Biol.,* vol. 6, pp. 413-431, 1964.

[14] J. Yin and J. S. McCaskill, "Replication of viruses in a growing plaque: a reaction-diffusion model," *Biophys. J.*, vol. 61, pp. 1540-1549, 1992.

[15] J. Yin, "Evolution of bacteriophage T7 in a growing plaque," *J. Bacteriol.*, vol. 175, pp. 1272-1277, 1993.

[16] Y. Lee and J. Yin, "Detection of evolving viruses," *Nat. Biotechnol.*, vol. 14, pp. 491-493, 1996.

[17] L. You and J. Yin, "Amplification and Spread of Viruses in a Growing Plaque," *J. Theor. Biol.*, vol. 200, pp. 365-373, 1999.

[18] E. L. Haseltine, V. Lam, J. Yin and J. B. Rawlings, "Image-guided modeling of virus growth and spread," *Bull. Math. Biol.*, vol. 70(6), pp. 1730-1748, 2008.

[19] J. Fort, "A comment on amplification and spread of viruses in a growing plaque," *J. Theor. Biol.*, vol. 214, pp. 515-518, 2002.

[20] V. Lam, K. A. Duca and J. Yin, "Arrested spread of vesicular stomatitis virus infections in vitro depends on interferon-mediated antiviral activity," *Biotechnol. Bioeng.*, vol. 90, no. 7, pp. 793-804, 2005.

[21] D. R. Amor and J. Fort, "Virus infection speeds: Theory versus experiment," *Phys. Rev. E*, vol. 82, no. 061905, 2010.

[22] C. Holland, "Glioblastoma multiforme: The terminator," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 97, pp. 6242-6244, 2000.

[23] K. Özduman, G. Wollmann, J. M. Piepmeier and A. N. van den Pol, "Systemic Vesicular Stomatitis Virus selectively destroys multifocal glioma and metastatic carcinoma in brain," *J. Neurosci.*, vol. 28, pp. 1882-1893, 2008.

[24] G. Wollmann, P. Tattersall and A. N. van den Pol, "Targeting human glioblastoma cells: comparison of nine viruses with oncolytic potential," *J. Virol.*, vol. 79, pp. 6005-6022, 2005.

[25] G. Wollmann, V. Rogulin, I. Simon, J. K. Rose and A. N. van den Pol, "Some attenuated variants of vesicular stomatitis virus show enhanced oncolytic activity against human glioblastoma cells relative to normal brain cells," *J. Virol.*, vol. 84, pp. 1563-1573, 2010.

[26] G. Wollmann, K. Ozduman and A. N. van den Pol, "Oncolytic virus therapy for glioblastoma multiforme: concepts and candidates," *Cancer J.*, vol. 18, pp. 69-81, 2012.

[27] R. L. Price and E. A. Chiocca, "Evolution of malignant glioma treatment: From chemotherapy to vaccines to viruses," *Neurosurgery*, vol. 61, pp. 74-83, 2014.

[28] "Oncolytic viruses," *Nature,* vol. 237, p. 486, 1972.

[29] E. Kelly and S. J. Russell, "History of oncolytic viruses: genesis to genetic engineering," *Mol. Ther.,* vol. 15, pp. 651-659, 2007.

[30] H. Ledford, "Cancer-fighting viruses near market," *Nature,* vol. 526, pp. 622-623, 2015.

[31] J. Altomonte, "Liver cancer: sensitizing hepatocellular carcinoma to oncolytic virus therapy," *Nature Rev. Gastroenterol. & Hepatol.,* vol. 15, pp. 8-10, 2018.

[32] S. J. Franks, H. M. Byrne, J. R. King, J. C. E. Underwood and C. E. Lewis, "Modelling the early growth of ductal carcinoma in situ of the breast," *J. Math. Biol.,* vol. 47, pp. 424-452, 2003.

[33] Y. Kuang, J. D. Nagy and S. E. Eikenberry, Introduction to mathematical oncology, Boca Raton: CRC Press, 2016.

[34] J. Fort and R. V. Solé, "Accelerated tumor invasion under non-isotropic cell dispersal in glioblastomas," *New J. Phys.,* vol. 15, p. 055001, 2013.

[35] M. Nowak and R. M. May, Virus dynamics: Mathematical principles of Immunology and Virology, Oxford: Oxford University Press, 2000, pp. 100-109.

[36] D. Wodarz, "Computational approaches to study oncolytic virus therapy: insights and challenges," *Gene Ther. Mol. Biol.,* vol. 8, pp. 137-146, 2004.

[37] D. Wodarz, "Gene therapy for killing p53-negative cancer cells: Use of replicating versus nonreplicating agents," *Hum. Gene Ther.,* vol. 14, pp. 153-159, 2004.

[38] D. Wodarz, Killer cell dynamics: mathematical and computational approaches to immunology, New York: Springer, 2006.

[39] D. Wodarz and N. Komarova, "Towards predictive computational models of oncolytic virus therapy: basis for experimental validation and model selection," *PLoS ONE,* vol. 4, no. e4271, 2009.

[40] D. Wodarz, C. N. Chan, B. Trinité, N. L. Komarova and D. N. Levy, "On the laws of virus spread through cell populations," *J. Virol.,* vol. 88, pp. 13240-13248, 2014.

[41] A. Friedman and Y. Tao, "Analysis of a model of a virus that replicates selectively un tumor cells," *J. Math. Biol.,* vol. 47, pp. 391-423, 2003.

[42] L. M. Wein, J. T. Wu and D. H. Kirn, "Validation and analysis of a mathematical model of a replication-competent oncolytic virus for cancer treatment: implications for virus design and delivery," *Cancer Res.,* vol. 63, pp. 1317-1324, 2003.

[43] W. Mok, T. Stylianopoulos, Y. Boucher and R. K. Jain, "Mathematical modeling of herpes simplex virus distribution in solid tumors: implications for cancer gene therapy," *Clin. Cancer Res.*, vol. 15, pp. 2352-2360, 2009.

[44] B. I. Camara, H. Mokrani and E. K. Afenya, "Mathematical modeling of glioma therapy using oncolytic viruses," *Math. Biosci. Eng.*, vol. 10, pp. 565-578, 2013.

[45] D. Wodarz, A. Hofacre, J. W. Lau, Z. Sun, H. Fan and N. L. Komarova, "Complex spatial dynamics of oncolytic viruses in vitro: mathematical and experimental approaches," *PLoS Comput. Biol.*, vol. 8, no. e1002547, 2012.

[46] D. Wodarz and N. Komarova, Dynamics of cancer: Mathematical foundations of oncology, New Jersey: World Scientific Publishing, 2014.

[47] J. Fort, T. Pujol and A.M. van der Linden, "Modelling the Neolithic transition in the Near East and Europe," *Am. Antiq.*, vol. 77, pp. 203-220, 2012.

[48] N. Isern, J. Zilhão, J. Fort and A. J. Ammerman, "Modeling the role of voyaging in the coastal spread of the Early Neolithic in the West Mediterranean," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 114, pp. 897-902, 2017.

[49] J. P. Bocquet-Appel and O. Bar-Yosef, The Neolithic demographic transition and its consequences, Berlin: Spinger, 2008.

[50] A. J. Ammerman and L. L. Cavalli-Sforza, "Measuring the rate of spread of early farming in Europe," *Man*, vol. 6, pp. 674-688, 1971.

[51] J. Fort and V. Méndez, "Time-delayed theory of the Neolithic transition in Europe," *Phys. Rev. Lett.*, vol. 82, pp. 867-870, 1999.

[52] A. J. Ammermann and L. L. Cavalli-Sforza, The Neolithic transition and the genetics of populations of Europe, Princeton: Princeton University Press, 1984.

[53] A. J. Ammerman and L. L. Cavalli-Sforza, "A population model for the diffusion of early farming in Europe," in *The explanation of culture change*, C. Renfrew, Ed., London: Duckworth, University of Pittsburgh Press, 1973, pp. 343-357.

[54] P. Menozzi, A. Piazza and L. L. Cavalli-Sforza, "Synthetic maps of human gene frequencies in Europeans," *Science*, vol. 201, pp. 786-792, 1978.

[55] L. L. Cavalli-Sforza, P. Menozzi and A. Piazza, "Demic expansions and human evolution," *Science*, vol. 259, pp. 639-646, 1993.

[56] C. A. Edmons, A. S. Lillie and L. L. Cavalli-Sforza, "Mutations arising in the wave front of an expanding population," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 101, pp. 975-979, 2004.

[57] G. Barbujani, "Genetic evidence for Prehistoric demographic changes in Europe," *Hum. Hered.,* vol. 76, pp. 133-141, 2013.

[58] R. Rasteiro and L. Chikhi, "Female and male perspectives on the Neolithic transition in Europe: Clues from ancient and modern genetic data," *PLoS ONE,* vol. 8, no. e60944, 2013.

[59] P. A. Underhill and T. Kivisild, "Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations," *Ann. Rev. Genet.,* vol. 41, pp. 539-564, 2007.

[60] O. Semino, G. Passarino, P. J. Oefner, A. A. Lin, S. Arbuzova, et al., "The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective," *Science,* vol. 290, pp. 1155-1159, 2000.

[61] B. Pakendorf and M. Stoneking, "Mitochondrial DNA and human evolution," *Annu. Rev. Genomics Hum. Genet.,* vol. 6, pp. 165-183, 2005.

[62] M. T. Seielstad, E. Minch, L. L. Cavalli-Sforza, "Genetic evidence for a higher female migration rate in humans," *Nature Genetics,* vol. 20, pp. 278-280, 1998.

[63] L. L. Cavalli-Sforza and A. Piazza, "Human genomic diversity in Europe: A summary of recent research and prospects for the future," *Eur. J. Hum. Genet.,* vol. 1, pp. 3-18, 1993.

[64] M. Richards, V. Macaulay, H. Bandelt and B. C. Sykes, "Phylogeography of mitochondrial DNA in western Europe," *Ann. Hum. Genet.,* vol. 62, pp. 241-260, 1998.

[65] A. Torroni, H. J. Bandelt, L. D'Urbano, P. Lahermo, P. Moral, et al., "mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe," *Am. J. Hum. Genet.,* vol. 62, p. 1137–1152, 1998.

[66] L. L. Cavalli-Sforza, P. Menozzi and A. Piazza, The history and geography of human genes, Princeton: Princeton Univ. Press, 1994.

[67] P. Bogucki, "The spread of early farming in Europe," *Am. Sci.,* vol. 84, pp. 242-253, 1996.

[68] M. Kaczanowska and J. K. Kozlowski, "Origins of the linear pottery complex and the Neolithic transition in Central Europe," in *The widening harvest. The Neolithic*

transition in Europe: looking back, looking forward, Boston: Archaeological Institute of America, A. J. Ammerman and P. Biagi, Eds., 2003, pp. 227-248.

[69] R. Clark, "The beginnings of agriculture in the sub-Alpine region: Some theoretical considerations," in *The Neolithisation of the Alpine region*, Brescia: Museo Civico di Scienze Naturali, P. Biagi, Ed., 1990, pp. 123-137.

[70] L. P. Louwe Kooijmans, "The gradual transition to farming in the Lower Rhine basin," in *Going over: the Mesolithic–Neolithic transition in north-west Europe*, Proceedings of the British Academy, vol. 144. London: British Academy, A. Whittle and V. Cummings, Eds., 2007, pp. 287-309.

[71] M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, et al., "Population genomics of Bronze Age Eurasia," *Nature*, vol. 522, pp. 167-172, 2015.

[72] L. L. Cavalli-Sforza, "Archaeology, genetics and language: reflecting on five decades of human genetics," in *Traces of ancestry: studies in honour of Colin Renfrew*, M. Jones, Ed., Cambridge: McDonald Institute for Archaeological Research, 2004, pp. 3-10.

[73] M. Richards, V. Macaulay, C. H. Hill, Á. Carradedo and A. Salas, "The archaeogenetics of the dispersals of the Bantu-speaking peoples," in *Traces of ancestry: studies in honour of Colin Renfrew*, M. Jones, Ed., Cambridge: McDonald Institute for Archaeological Research, 2004, pp. 75-87.

[74] J. B. Pererira, M. D. Costa, D. Vieira, M. Pala, L. Bamford, et al., "Reconciling evidence from ancient and contemporary genomes: a major source for the European Neolithic within Mediterranean Europe," *Proc. Roy. Soc. B*, vol. 284, no. 20161976, 2017.

[75] C. Gamba, E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes, et al., "Genome flux and stasis in a five millennium transect of European prehistory," *Nat. Commun.*, vol. 5, no. 5257, 2014.

[76] W. Haak, P. Forster, B. Bramanti, S. Matsumura, G. Brandt, et al., "Ancient DNA from the first European farmers in 7,500 year old Neolithic sites," *Science*, vol. 310, pp. 1016-1018, 2005.

[77] R. Pinhasi, M. G. Thomas, M. Hofreiter, M. Currat and J. Burger, "The genetic history of Europeans," *Trends in Genetics*, vol. 28, pp. 496-505, 2012.

[78] G. Brandt, W. Haak, C. Adler, C. Roth, A. Szécsényi-Nagy, et al., "Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity," *Science*, vol. 342, pp. 257-261, 2013.

[79] M. Rasmussen, Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, et al., "Ancient human genome sequence of an extinct Palaeo-Eskimo," *Nature,* vol. 463, pp. 757-762, 2010.

[80] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, et al., "Genome-wide patterns of selection in 230 ancient Eurasians," *Nature,* vol. 528, pp. 499-503, 2015.

[81] I. Lazaridis, D. Nadel, G. Rollefson, D. Merrett, N. Rohland, et al., "Genomic insights into the origin of farming in the ancient Near East," *Nature,* vol. 536, pp. 419-424, 2016.

[82] A. Mittnik, C.-C. Wang, S. Pfrengle, M. Daubaras, G. Zarina, et al., "The genetic prehistory of the Baltic Sea region," *Nature Commun.,* vol. 9, no. 442, 2018.

[83] W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, et al., "Massive migration from the steppe was a source for Indo-European languages in Europe," *Nature,* vol. 522, p. 207–211, 2015.

[84] I. Olalde, S. Brace, M. Allentoft, I. Armit, K. Kristiansen, et al., "The Beaker phenomenon and the genomic transformation of northwest Europe," *Nature,* vol. 555, p. 190–196, 2018.

[85] F. De Angelis, G. Scorrano, C. Martínez-Labarga, G. Scano, F. Macciardi and O. Rickards, "Mitochondrial variability in the Mediterranean area: A complex stage for human migratins," *Ann. Hum. Genet.,* vol. 45, pp. 5-19, 2018.

[86] I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, et al., "The genomic history of southeastern Europe," *Nature,* vol. 555, pp. 197-203, 2018.

[87] S. Brace, Y. Diekmann, T. J. Booth, Z. Faltyskova, N. Rohland, et al., "Population replacement in Early Neolithic Britain," *bioRxiv,* http://dx.doi.org/10.1101/267443.

[88] G. Barbujani, R. Sokal and N. Oden, "Indo-European origins: A computer-simulation test of five hypotheses," *Am. J. Phys. Anthropol.,* vol. 96, pp. 109-132, 1995.

[89] M. Currat and L. Excoffier, "The effect of the Neolithic expansion on European molecular diversity," *Proc. R. Soc. B,* vol. 272, pp. 679-688, 2005.

[90] P. Sjödin and O. François, "Wave-of-advance models of the diffusion of the Y chromosome haplogroup R1b1b2 in Europe," *PLoS ONE,* vol. 6, no. e21592, 2011.

[91] P. Paschou, P. Drineas, E. Yannaki, A. Razou, K. Kanaki, et al., "Maritime route of colonization of Europe," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 111, pp. 9211-9216, 2014.

[92] K. Aoki, M. Shida and N. Shigesada, "Travelling wave solutions for the spread if farmers into a region occupied by hunter-gatherers," *Theor. Popul. Biol.,* vol. 50, pp. 1-17, 1996.

[93] L. L. Cavalli-Sforza and M. W. Feldman, Cultural transmission and evolution: A quantitative approach, Princeton: Princeton University Press, 1981.

[94] J. Fort, "Vertical cultural transmission effects on demic front propagation: Theory and application to the Neolithic transition in Europe," *Phys. Rev. E,* vol. 83, p. 056124, 2011.

[95] J. Fort, "Demic and cultural diffusion propagated the Neolithic transition across different regions of Europe," *J. R. Soc. Interface,* vol. 12, no. 20150166, 2015.

[96] R. A. Fisher, "The wave of advance of advantageous genes," *Ann. Hum. Genet.,* vol. 7, pp. 353-369, 1937.

[97] D. Frenkel and B. Smit, Understanding molecular simulation: from algorithms to applications, San Diego: Academic Press, 1996.

[98] A. Einstein, Investigations on the theory of the Brownian movement, New York: Dover, 1956.

[99] A. N. Kolmogorov, I. G. Petrovsky and N. S. Piskunov, "Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique," *Moscow Univ. Math. Bull.*, vol. 1, pp. 1-26, 1937.

[100] R. B. Bird, W. E. Stewart and E. N. Lightfoot, Transport Phenomena, New York: John Wiley & Sons, 1976.

[101] J. Crank, The mathematics of diffusion, 2nd ed., Oxford: Oxford University Press, 1975.

[102] H. C. Berg, Random walks in Biology, Princeton: Princeton Univ. Press, 1993.

[103] B. S. Bokshtein, M. I. Mendelev and D. J. Srolovitz, Thermodynamics and Kinetics in Materials Science, Oxford: Oxford University Press, 2005, pp. 165-175.

[104] J. Murray, Mathematical Biology, 3rd ed., New York: Springer-Verlag, 2002.

[105] G. F. Gause, The struggle for existence, Baltimore, Maryland: The Williams & Wilkins company, 1934.

[106] N. J. Gotelli, A primer of ecology, 4th ed., Sunderland, Massachusetts: Sinauer Associates Inc., 2008.

[107] J. G. Skellam, "Random dispersal in theoretical populations," *Biometrika,* vol. 38, pp. 196-218, 1951.

[108] N. Shigesada and K. Kawasaki, Biological invasions: theory and practice, Oxford: Oxford Univ. Press, 1997.

[109] L. L. Cavalli-Sforza, "Recollections of Whittingehame Lodge," *Theor. Popul. Biol.,* vol. 38, pp. 301-305, 1990.

[110] J. Fort, T. Pujol and L. Cavalli-Sforza, "Palaeolithic populations and waves of advance," *Cambridge Archaeol. J.,* vol. 14, pp. 53-61, 2004.

[111] M. Hamilton and B. Buchanan, "Spatial gradients in Clovis-age radiocarbon dates across North America suggest rapid colonization from the north," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 104, pp. 15625-15630, 2007.

[112] V. Ortega-Cejas, J. Fort and V. Méndez, "Role of the delay time in the modelling of biological range expansions," *Ecology,* vol. 85, pp. 258-264, 2004.

[113] S. R. Dunbar, "Travelling wave solutions of diffusive Lotka-Volterra equations," *J. Math. Biol.,* vol. 17, pp. 11-32, 1983.

[114] G. J. Bauer, J. S. McCaskill and H. Otten, "Traveling waves of in vitro evolving RNA," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 86, pp. 7937-7941, 1989.

[115] V. Ortega-Cejas, J. Fort, V. Méndez and D. Campos, "Approximate solution to the speed of spreading viruses," *Phys. Rev. E,* vol. 69, no. 031909, 2004.

[116] J. Fort, J. Pérez-Losada, E. Ubeda and F. J. García, "Fronts with continuous waiting-time distributions & virus infections," *Phys. Rev. E,* vol. 73, no. 021907, 2006.

[117] N. Isern and J. Fort, "Time-delayed reaction-diffusion fronts," *Phys. Rev. E,* vol. 80, no. 057103, 2009.

[118] S. J. Russell, K. W. Peng and J. C. Bell, "Oncolytic virotherapy," *Nat. Biotechnol.,* vol. 30, pp. 658-670, 2012.

[119] R. H. I. Andtbacka, H. L. Kaufman, F. Collichio, T. Amatruda, N. Senzer, et. al., "Talimogene laherparepvec improves durable response rate in patients with advanced melanoma," *J. Clin. Oncol.,* vol. 33, pp. 2780-2788, 2015.

[120] H. L. Kaufman, F. J. Kohlhapp and A. Zloza, "Oncolytic viruses: a new class of immunotherapy drugs," *Nat. Rev. Drug Discov.,* vol. 14, pp. 642-662, 2015.

[121] I. A. Rodriguez-Brenes, A. Hofacre, H. Fan and D. Wodarz, "Complex dynamics of virus spread from low infection multiplicities: implications for the spread of oncolytic viruses," *PLoS Comput. Biol.,* vol. 13, no. e1005241, 2017.

[122] B. Vogelstein and K. Kinzler, "Cancer genes and the pathways they control," *Nat. Med.,* vol. 10, pp. 789-799, 2004.

[123] R. A. Weinberg, The Biology of Cancer, New York: Garland Science, 2007.

[124] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumors," *Nature,* vol. 490, pp. 61-70, 2012.

[125] R. Meza, J. Jeon, S. H. Moolgavkar and E. G. Luebeck, "Age-specific incidence of cancer: Phases, transitions, and biological implications," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 105, p. 16284–16289, 2008.

[126] A. S. Novozhilov, F. S. Berezovskaya, E. V. Koonin and G. P. Karev, "Mathematical modeling of tumor therapy with oncolytic viruses: Regimes with complete tumor elimination within the framework of deterministic models," *Biology Direct,* vol. 1, no. 6, 2006.

[127] G. P. Karev, A. S. Novozhilov and E. V. Koonin, "Mathematical modeling of tumor therapy with oncolytic viruses: effects of parametric heterogeneity on cell dynamics," *Biology Direct,* vol. 1, no. 30, 2006.

[128] N. L. Komarova and D. Wodarz, "ODE models for oncolytic virus dynamics," *J. Theor. Biol.,* vol. 263, pp. 530-543, 2010.

[129] R. M. May and R. M. Anderson, "Population biology of infectious diseases: Part II," *Nature,* vol. 280, pp. 455-461, 1979.

[130] E. Beretta and Y. Kuang, "Modeling and analysis of a marine bacteriophage infection," *Math. Biosci.,* vol. 149, pp. 57-76, 1998.

[131] V. Childe, The dawn of European civilisation, London: Routledge, 1925.

[132] V. G. Childe, What happened in history, Harmondsworth: Penguin Books, 1942.

[133] M. S. Edmonson, "Neolithic diffusion rates," *Curr. Anthropol.,* vol. 2, pp. 71-102, 1961.

[134] J. G. D. Clark, "Radiocarbon dating and the spread of the farming economy," *Antiquity,* vol. 39, pp. 45-48, 1965.

[135] E. Neustupný, "Absolute chronology of the Neolithic and Aeneolithic periods in central and south-eastern Europe," *Slovenská archeológia,* vol. 16, pp. 19-56, 1968.

[136] P. Ucko and G. W. Dimbleby, The domestication and exploitation of plants and animals, London: Duckworth, 1969.

[137] K. Davison, P. Dolukhanov, G. R. Sarson and A. Shukurov, "The role of waterways in the spread of the Neolithic," *J. Archaeol. Sci.*, vol. 33, pp. 641-652, 2006.

[138] J. Bernabeu Aubán, C. M. Barton, S. Pardo Godó and S. M. Bergin, "Modeling initial Neolithic dispersal. The first agricultural groups in west Mediterranean," *Ecol. Modell.*, vol. 307, pp. 22-31, 2015.

[139] J. Fort, J. Pérez-Losada and N. Isern, "Fronts from integrodifference equations and persistence effects on the Neolithic transition," *Phys. Rev. E,* vol. 76, no. 031913, 2007.

[140] M. A. Zeder, "Domestication and early agriculture in the Mediterranean basin: Origins, diffusion, impact.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 1015, pp. 11597-11604, 2008.

[141] K. Kujit and N. Going-Moris, "Foraging, farming, and social complexity in the pre-pottery of the Neolithic of the southern Levant: A review and synthesis," *J. World Prehist.*, vol. 16, pp. 361-440, 2002.

[142] L. L. Cavalli-Sforza, African pygmies, Orlando: Academic Press, 1986, pp. 361-426.

[143] R. A. Bentley, R. H. Layton and J. Tehrani, "Kinship, marriage, and the genetics of past human dispersals," *Hum. Biol.*, vol. 81, pp. 159-179, 2009.

[144] L. You and J. Yin, "Amplification and spread of viruses in a growing plaque," *J. Theor. Biol.*, vol. 200, pp. 365-373, 1999.

[145] D. R. Amor and J. Fort, "Cohabitation reaction-diffusion model for virus focal infections," *Physica A,* vol. 416, pp. 611-619, 2014.

[146] S. A. Gourley and Y. Kuang, "A delay reaction-diffusion model of the spread of bacteriophage infection," *SIAM J. Appl. Math.*, vol. 65, pp. 550-566, 2005.

[147] D. A. Jones, H. L. Smith, H. R. Thieme and G. Röst, "On spread of phage infection of bacteria in a petri dish," *SIAM J. Appl. Math.*, vol. 72, pp. 670-688, 2012.

[148] H. W. Ackermann, "La classification des phages caudés des entérobacteries," *Pathol. Biol.*, vol. 24, pp. 359-380, 1976.

[149] H. Fricke, "A mathematical treatment of the electric conductivity and capacity of disperse systems I. The electric conductivity of a suspension of homogeneous spheroids," *Phys. Rev.*, vol. 24, pp. 575-587, 1924.

[150] T. D. Brock and M. T. Madigan, Biology of Microorganisms, 6th ed., Englewood Cliffs: Prentice-Hall, 1991, p. 40.

[151] B. R. Ware, T. Raj, W. H. Flygare, J. A. Lesnaw and M. E. Reichmann, "Molecular weights of Vesicular Stomatitis Virus and its defective particles by laser light-scattering spectroscopy," *J. Virol.,* vol. 11, pp. 141-145, 1973.

[152] A. M. Stein, D. A. Vader, T. S. Deisboeck, E. A. Chiocca, L. M. Sander and D. A. Weitz, "Directionality of glioblastoma invasion in a 3D in vitro experiment," *http://arxiv.org/pdf/q-bio/0610031.pdf (accessed 30 September 2012),* 2006.

[153] A. M. Stein, T. Demuth, D. Mobley, M. Berens and L. M. Sander, "A mathematical model of glioblastoma tumor spheroid invasion in a three-dimensional in vitro experiment," *Biophys. J.,* vol. 92, pp. 356-365, 2007.

[154] K. Shishido, A. Watarai, S. Naito and T. Ando, "Action of bleomycin on the bacteriophate T7 infection," *J. Antibiot.,* vol. 28, pp. 676-680, 1975.

[155] G. Wollmann, M. D. Robek and A. N. van den Pol, "Variable deficiencies in the interferon response enhance susceptibility to vesicular stomatitis virus oncolytic actions in glioblastoma cells but not in normal human glial cells," *J. Virol.,* vol. 81, pp. 1479-1491, 2007.

[156] H. Ikeda, R. J. de Boer, K. Sato, S. Morita, N. Misawa, et al., "Improving the estimation of the death rate of infected cells from time course data during the acute phase of virus infections: application to acute HIV-1 infection in a humanized mouse model," *Theor. Biol. Med. Model.,* vol. 11, pp. 22-35, 2014.

[157] S. Arnold, M. Siemann, K. Scharnweber, M. Werner, S. Bauman and S. Reuss, "Kinetic modeling and simulation of in vitro transcription by phage T7 RNA polymerase," *Biotechnol. Bioeng.,* vol. 72, pp. 548-561, 2001.

[158] A. N. van den Pol and J. N. Davis, "Highly attenuated recombinant vesicular stomatitis virus VSV-12'GFP displays immunogenic and oncolytic activity," *J. Virol.,* vol. 87, pp. 1019-1034, 2013.

[159] E. L. Haseltine, V. Lam, J. Yin and J. B. Rawlings, "Image-guided modeling of virus growth and spread," *Bull. Math. Biol.,* vol. 70, pp. 1730-1748, 2008.

[160] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnell, Molecular cell Biology, 4th ed., New York: W. H. Freeman, 2000.

[161] R. Rockne, J. K. Rockhill, M. Mrugala, A. M. Spence, I. Kalet, et al., "Predicting the efficacy of radiotherapy in individual glioblastoma patients in vivo: a mathematical modeling approach," *Phys. Med. Biol.,* vol. 55, pp. 3271-3285, 2010.

[162] S. E. Eikenberry, T. Sankar, M. C. Preul, E. J. Kostelich, C. J. Thalhauser and T. Kuang, "Virtual glioblastoma: Growth, migration and treatment in a three-dimensional mathematical model," *Cell Prolif.,* vol. 42, pp. 511-528, 2009.

[163] A. Friedman, J. P. Tian, G. Fulci, E. A. Chiocca and J. Wang, "Glioma virotherapy: Effects of innate immune suppression and increased viral replication capacity," *Cancer Res.,* vol. 66, pp. 2314-2319, 2006.

[164] S. Manley, J. M. Gillette, G. H. Patterson, H. Shroff, H. F. Hess, et al., "High-density mapping of single-molecule trajectories with photoactivated localization microscopy," *Nat. Methods,* vol. 5, pp. 155 - 157, 2008.

[165] E. Fernández, A. Pérez-Pérez, C. Gamba, E. Prats, P. Cuesta, et al., "Ancient DNA analysis of 8,000 B.C. Near Eastern farmers supports an early Neolithic pioneer maritime colonization of mainland Europe through Cyprus and the Aegean Islands," *PLoS Genet.,* vol. 10, no. e1004401, 2014.

[166] C. Gamba, E. Fernández, M. Tirado, M. Deguilloux, M. Pemonge, et al., "Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers," *Mol. Ecol.,* vol. 21, pp. 45-56, 2012.

[167] M. Hervella, N. Izagirre, S. Alonso, R. Fregel, A. Alonso, et al., "Ancient DNA from hunter-gatherer and farmers groups from Northern Spain supports a random dispersion model for the Neolithic expansion into Europe," *PLoS ONE,* vol. 7, no. e34417, 2012.

[168] H. Malmström, M. Gilbert, M. Thomas, M. Brandström, J. Stora, et al., "Ancient DNA Reveals Lack of Continuity between Neolithic Hunter-Gatherers and Contemporary Scandinavians," *Curr. Biol.,* vol. 19, pp. 1758-1762, 2009.

[169] W. Haak, O. Balanovsky, J. J. Sánchez, S. Koshel, V. Zaporozhchenko, et al., "Ancient DNA from European early Neolithic farmers reveals their Near Eastern affinities," *PLoS Biol.,* vol. 8, no. e1000536, 2010.

[170] Ç. Çilingiroglu, "The concept of 'Neolithic package': considering its meaning and applicability," *Documenta Praehistorica,* vol. 32, pp. 1-13, 2005.

[171] Z. Hofmanová, S. Kreutzer, G. Hellenthal, C. Sell, Y. Diekmann, et al., "Early farmers from across Europe directly descended from Neolithic Aegeans," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 113, pp. 6886-6891, 2016.

[172] A. Szécsény-Nagy, G. Brandt, W. Haak, V. Keerl, J. Jakucs, et al., "Tracing the genetic origin of Europe's first farmers reveals insights into their social organizations," *Proc. R. Soc. B,* vol. 282, no. 20150339, 2015.

[173] I. Olalde, H. Schroeder, M. Sandoval-Velasco, L. Vinner, I. Lobón, et al., "A common genetic origin for early farmers from Mediterranean cardial and central european LBK cultures," *Mol. Biol. Evol.,* vol. 32, pp. 3132-3142, 2015.

[174] H. Chandler, B. Sykes and J. Zilhão, "Using ancient DNA to examine genetic continuity at the Mesolithic-Neolithic transition in Portugal," in *Actas dell III Congreso del Neolítico en la Península Ibérica,* vol. 1, P. Arias, R. Ontañón and C. García-Moncó, Eds., Santander: Monografías del Instituto internacional de Investigaciones Prehistóricas de Cantabria, 2005, pp. 781-786.

[175] H. Chandler, Using Ancient DNA to Link Culture and Biology in Human Populations, Oxford: University of Oxford, 2003.

[176] H. Malmström, A. Linderholm, P. Skoglund, J. Stora, P. Sjödin, et al., "Ancient mitochondrial DNA from the northern fringe of the Neolithic farming expansion in Europe sheds light on the dispersion process," *Philos. Trans. R. Soc. Lond., B, Biol. Sci.,* vol. 370, no. 20130373, 2015.

[177] P. Skoglund, H. Malmström, A. Omrak, M. Raghavan, C. Valdiosera, et al., "Genomic diversity and admixture differs for stone-age scandinavian foragers and farmers," *Science,* vol. 344, pp. 747-750, 2014.

[178] J. Stauder, The Majangir. Ecology and Society of a Southwest Ethiopian People, London: Cambridge University Press, 1971.

[179] J. Fort, J. Pérez-Losada, J. J. Suñol, L. Escoda and J. M. Massaneda, "Integro-difference equations for interacting species and the Neolithic transition," *New J. Phys.,* vol. 10, no. 43045, 2008.

[180] J. Steele, J. M. Adams and T. Sluckin, "Modeling Paleoindian dispersals," *World Archaeol.,* vol. 30, pp. 286-305, 1998.

[181] J. Alroy, "A multispecies overkill simulation of the end-Pleistocene megafaunal mass extinction," *Science,* vol. 292, pp. 1893-1896, 2001.

[182] J. Fort, D. Jana and J. M. Humet, "Multidelayed random walks: Theory and application to the neolithic transition in Europe," *Phys. Rev. E,* vol. 70, no. 031913, 2004.

[183] N. Isern, J. Fort and J. Pérez-Losada, "Realistic dispersion kernels applied to cohabitation reaction–dispersion equations," *J. Stat. Mech. Theor. Exp.,* vol. 2008, no. 10, p. P10012, 2008.

[184] J. P. Birdsell, "Some population problems involving Pleistocene man," *Cold Spring Harb. Symp. Quant. Biol.,* vol. 22, pp. 47-69, 1957.

[185] D. F. Roberts, "Genetic effects of population size reduction," *Nature,* vol. 220, pp. 1084-1088, 1968.

[186] A. J. Lotka, Elements of Mathematical Biology, New York: Dover, 1956, pp. 64-69.

[187] L. L. Cavalli-Sforza, "The distribution of migration distances: models and applications to genetics," in *Les Deplacements Humains*, L. L. Cavalli-Sforza and J. Sutter, Eds., Monaco: Editions Sciences Humaines, 1962, pp. 139-158.

[188] J. Fort and T. Pujol, "Progress in front propagation research," *Rep. Prog. Phys.,* vol. 71, no. 086001, 2008.

[189] U. Ebert and W. van Saarloos, "Front propagation into unstable states: Universal algebraic convergence towards uniformly translating pulled fronts," *Physica D,* vol. 146, pp. 1-99, 2000.

[190] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, Numerical Recipes: The Art of Scientific Computing, New York: Cambridge University Press, 2007.

[191] M. G. Weinbauer, "Ecology of prokaryotic viruses," *FEMS Microbiol. Rev.,* vol. 28, pp. 127-181, 2004.

[192] J. Yin, "Spatially resolved evolution of viruses," *Ann. N. Y. Acad. Sci.,* vol. 745, pp. 399-408, 1994.

[193] M. de Paepe and F. Taddei, "Viruses' life history: Towards a mechanistic basis of a trade-off between survival and reproduction among phages," *PLoS Biol.,* vol. 4, no. e193, 2006.

[194] C. E. Zobell and A. B. Cobet, "Growth, reproduction, and death rates of Escherichia coli at increased hydrostatic pressures," *J. Bacteriol.,* vol. 84, pp. 1228-36, 1962.

[195] J. Fort and V. Méndez, "Reaction-diffusion waves of advance in the transition to agricultural economics," *Phys. Rev. E,* vol. 60, pp. 5894-5901, 1999.

[196] A. I. Freeman, Z. Zakay-Rones, J. M. Gomori, E. Linetsky, L. Rasooly, et al., "Phase I/II trial of intravenous NDV-HUJ oncolytic virus in recurrent glioblastoma multiforme," *Mol. Ther.,* vol. 13, pp. 221-228, 2006.

[197] F. Brauer and C. Castillo-Chavez, Mathematical models in population biology and epidemiology, New York: Springer, 2001, pp. 123-125.

[198] T. L. Stepien, E. M. Rutter and Y. Kuang, "A data-motivated density-dependent diffusion model of in vitro glioblastoma growth," *Math. Biosci. Eng.,* vol. 12, pp. 1157-1172, 2015.

[199] C. A. Koks, S. De Vleeschouwer, N. Graf and S. W. Van Gool, "Immune suppression during oncolytic virotherapy for high-grade glioma; yes or no?," *J. Cancer,* vol. 6, pp. 203-217, 2015.

[200] D. J. Mahoney, D. F. Stojdl and G. Laird, "Virus therapy for cancer," *Sci. Am.,* vol. 311, pp. 54-59, 2014.

[201] R. Pinhasi, J. Fort and A. J. Ammerman, "Tracing the origin and spread of agriculture in Europe," *PLoS Biol.,* vol. 3, no. e410, 2005.

[202] N. J. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, et al., "Ancient admixture in human history," *Genetics,* vol. 192, pp. 1065-1093, 2012.

[203] B. Bramanti, M. G. Thomas, W. Haak, M. Unterlaender, P. Jores, et al., "Genetic discontinuity between local hunter-gatherers and central Europe's first farmers," *Science,* vol. 326, pp. 137-140, 2009.

[204] Q. Fu, C. Posth, M. Hajdinjak, M. Petr, S. Mallick, et al., "The genetic history of Ice Age Europe," *Nature,* vol. 534, pp. 200-205, 2016.

[205] Q. Atkinson, "Phonemic diversity supports a serial founder effect model of language expansion from Africa," *Science,* vol. 332, pp. 346-349, 2011.

[206] M. Sampietro, O. Lao, D. Caramelli, M. Lari, R. Pou, et al., "Palaeogenetic evidence supports a dual model of Neolithic spreading into Europe," *Proc. R. Soc. B,* vol. 274, pp. 2161-2167, 2007.

[207] M. Rivollat, S. Rottier, C. Couture, M.-H. Pemonge, F. Mendisco, et al., "Investigating mitochondrial DNA relationships in Neolithic Western Europe through serial coalescent simulations," *Eur. J. Hum. Genet.,* vol. 25, no. 3, pp. 388-392, 2017.

[208] L. Cronk, "From hunters to herders: Subsistence change as a reproductive strategy among the Mukogodo," *Curr. Anthropol.,* vol. 30, pp. 224-234, 1989.

[209] J. Early and T. Headland, Population dynamics of a Philippine rain forest people: The San Ildefonso Agta, Gainesville: University of Florida Press, 1998.

[210] L. Excoffier, G. Laval and S. Schneider, "Arlequin ver 3.0: An integrated software package for population genetics data analysis," *Evol. Bioinform. Online,* vol. 1, pp. 47-50, 2005.

[211] F. Tajima, "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism," *Genetics,* vol. 123, no. 3, pp. 585-595, 1989.

[212] Y. X. Fu, "Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection," *Genetics,* vol. 147, no. 2, pp. 915-925, 1997.

[213] L. Excoffier and S. Schneider, "Why hunter-gatherer populations do not show signs of pleistocene demographic expansions," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 96, no. 19, pp. 10597-10602, 1999.

[214] L. Pereira, I. Dupanloup, Z. Rosser, M. Jobling and G. Barbujani, "Y-chromosome mismatch distributions in Europe," *Mol. Biol. Evol.,* vol. 18, no. 7, pp. 1259-1271, 2001.

[215] R. M. Andrews, I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull and N. Howell, "Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA," *Nat. Genet.,* vol. 23, no. 2, p. 147, 1999.

[216] M. Nei, Molecular evolutionary genetics, New York: Columbia University Press, 1987.

[217] J. Elsner, M. Hofreiter, J. Schibler and A. Schlumbaum, "Ancient mtDNA diversity reveals specific population development of wild horses in Switzerland after the Last Glacial Maximum," *PLoS One,* vol. 12, no. 5, p. e0177458, 2017.

[218] A. R. Rogers and H. Harpending, "Population growth makes waves in the distribution of pairwise genetic differences," *Mol. Biol. Evol.,* vol. 9, no. 3, pp. 552-569, 1992.

[219] N. Ray, M. Currat and L. Excoffier, "Intra-deme molecular diversity in spatially expanding populations," *Mol. Biol. Evol.,* vol. 20, no. 1, pp. 76-86, 2003.

[220] L. Excoffier, "Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model," *Mol. Ecol.,* vol. 13, no. 4, pp. 853-864, 2004.

[221] S. Ramachandran, et al., "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 102, no. 44, pp. 15942-15947, 2005.

[222] M. Slatkin, "Isolation by distance in equilibrium and non-equilibrium populations," *Evolution,* vol. 47, no. 1, pp. 264-279, 1993.

[223] M. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer Res. ,* vol. 27, no. 2, pp. 209-220, 1967.

[224] P. E. Smouse, J. C. Long and R. R. Sokal, "Multiple regression and correlation extensions of the mantel test of matrix correspondence," *Syst. Zool.,* vol. 35, no. 4, pp. 627-632, 1986.

[225] F. Messina, G. Scano, I. Contini, C. Martínez-Labarga, G. F. De Stefano and O. Rickards, "Linking between genetic structure and geographical distance: Study of the

maternal gene pool in the Ethiopian population," *Ann. Hum. Biol.,* vol. 44, no. 1, pp. 53-69, 2017.

[226] P. Legendre and M. J. Fortin, "Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data," *Mol. Ecol. Resour.,* vol. 10, no. 5, pp. 831-844, 2010.

[227] Ø. Hammer, D. A. T. Harper and P. D. Ryan, "PAST: paleontological statistics software package for education and data analysis," *Palaeontol. Electron.,* vol. 4, no. 1, pp. 1-9, 2001.

[228] M. Slatkin and R. R. Hudson, "Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations," *Genetics,* vol. 129, no. 2, pp. 555-562, 1991.

[229] H. J. Bandelt, P. Forster and A. Röhl, "Median-joining networks for inferring intraspecific phylogenies," *Mol. Biol. Evol.,* vol. 16, no. 1, pp. 37-48, 1999.

[230] S. Barnabas, Y. Shouche and C. G. Suresh, "High-resolution mtDNA studies of the Indian population: implications for palaeolithic settlement of the Indian subcontinent," *Ann. Hum. Genet.,* vol. 70, pp. 42-58, 2006.

[231] A. J. Drummond, A. Rambaut, B. Shapiro and O. G. Pybus, "Bayesian coalescent inference of past population dynamics from molecular sequences," *Mol. Biol. Evol.,* vol. 22, no. 5, pp. 1185-1192, 2005.

[232] R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C. H. Wu, et al., "BEAST 2: a software platform for Bayesian evolutionary analysis," *PLoS Comput. Biol.,* vol. 10, no. 4, p. e1003537, 2014.

[233] A. Rambaut, M. Suchard and A. Drummond, "Tracer," 2013. [Online]. Available: http://tree.bio.ed.ac.uk/software/tracer/.

[234] P. Soares, L. Ermini, N. Thomson, M. Mormina, T. Rito, et al., "Correcting for purifying selection: An improved human mitochondrial molecular clock," *Am. J. Hum. Genet.,* vol. 84, no. 6, pp. 740-759, 2009.

[235] K. R. Zenger, B. J. Richardson and A. Vachot-Griffin, "A rapid population expansion retains genetic diversity within European rabbits in Australia," *Mol. Ecol.,* vol. 12, no. 3, pp. 789-794, 2003.

[236] L. Xu, H. Xue, M. Song, Q. Zhao, J. Dong, et al., "Variation of genetic diversity in a rapidly expanding population of the greater long-tailed hamster (Tscherskia triton) as revealed by microsatellites," *PLoS ONE,* vol. 8, no. 1, p. e54171, 2013.

[237] S.M. Murphy, J.J. Cox, J.D. Clark, B.C. Augustine, J.T. Hast, et al., "Rapid growth and genetic diversity retention in an isolated reintroduced black bear population in the central appalachians," *Wildl. Soc. Bull.,* vol. 79, no. 5, pp. 807-818, 2015.

[238] V. Coia, G. Cipollini, P. Anagnostou, F. Maixner, C. Battaggia, et al., "Whole mitochondrial DNA sequencing in Alpine populations and the genetic history of the Neolithic Tyrolean Iceman," *Sci. Rep.,* vol. 6, p. 18932, 2016.

[239] G. Eriksson, A. Linderholm, E. Fornander, M. Kanstrup, P. Schoultz, et al., "Same island, different diet: Cultural evolution of food practice on Öland, Sweden, from the Mesolithic to the Roman Period," *J. Anthropol. Archaeol.,* vol. 27, no. 4, pp. 520-543, 2008.

[240] M. Malmer, The Neolithic of south Sweden: TRB, GRK, and STR, Stockholm: Royal Swedish Academy of Letters History and Antiquities, 2002.

[241] P. Legendre and L. Legendre, Numerical Ecology, 3rd ed., Amsterdam: Elsevier Science, 2012.

[242] F. Messina, A. Finocchio, N. Akar, A. Loutradis, E.I. Michalodimitrakis, et al., "Spatially explicit models to investigate geographic patterns in the distribution of forensic STRs: Application to the North-Eastern Mediterranean," *PLoS ONE,* vol. 11, no. 11, p. e0167065, 2016.

[243] M. S. Rosenberg and C. D. Anderson, "PASSaGE: Pattern analysis, spatial statistics and geographic exegesis. Version 2," *Meth. Ecol. Evol.,* vol. 2, no. 3, pp. 229-232, 2011.

[244] N. L. Oden, "Assessing the significance of a spatial correlogram," *Geogr. Anal.,* vol. 16, pp. 1-16, 1984.

[245] J. Zilhao, "Radiocarbon evidence for maritime pioneer colonization at the origins of farming in west Mediterranean Europe," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 98, pp. 14180-14185, 2001.

[246] J. Fort, T. Pujol and M. Vander-Linden, "Modelling the Neolithic transition in the Near East and Europe," *Am. Antiq.,* vol. 77, pp. 203-220, 2012.

[247] N. Isern, J. Fort, A. Carvalho, J. Gibaja and J. Ibañez, "The Neolithic transition in the Iberian Peninsula: data analysis and modelling," *J. Archaeol. Method Th.,* vol. 21, pp. 447-460, 2014.

[248] H. M. Wobst, "Boundary conditions for paleolithic social systems: A simulation approach," *Am. Antiq.,* vol. 39, no. 2, pp. 147-178, 1974.

[249] Y. Raviv and N. Intrator, "Bootstrapping with noise: An effective regularization technique," *Connection Science,* vol. 8, no. 3-4, pp. 355-372, 1996.

[250] B. Efron, The jackknife, the bootstrap and other resampling plans, Montpelier, Vermont: The Society for Industrial and Applied Mathematics (SIAM), 1982.

[251] P. Dixon, "The bootstrap and the jackknife: describing the precision of ecological indices," in *Design and analysis of ecological experiments,* S. M. Scheiner and J. Gurevitch, Eds., New York: Chapman and Hall, 1993, pp. 267-288.

[252] Q. Fu, A. Mittnik, P.L.F. Johnson, K. Bos, M. Lari, et al., "A revised timescale for human evolution based on ancient mitochondrial genomes," *Curr. Biol.,* vol. 23, no. 7, pp. 553-559, 2013.

[253] M. Lacan, C. Keyser, F.X. Ricaut, N. Brucato, F. Duranthon, et al. , "Ancient DNA reveals male diffusion through the Neolithic Mediterranean route," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 108, pp. 9788-9791, 2011.

[254] P. Brotherton, W. Haak, J. Templeton, G. Brandt, J. Soubrier, et al., "Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans," *Nat. Comm.,* vol. 4, p. 1764, 2013.

[255] Q. Fu, P. Rudan, S. Pääbo and J. Krause, "Complete mitochondrial genomes reveal neolithic expansion into Europe," *PLoS One,* vol. 7, no. 3, p. e32473, 2012.

[256] M. Demerec and U. Fano, "Bacteriophage-Resistant Mutants in Escherichia Coli," *Genetics,* vol. 30, pp. 119-136, 1945.

[257] V. M. Narasimhan, N. Paerson, P. Moorjani, I. Lazardis, M. Lipson, et al., "The genomic formation of south and central Asia," *bioRxiv,* http://dx.doi.org/10.1101/292581.

[258] M. Hervella, M. Rotea, N. Izagirre, M. Constantinescu, S. Alonso, et al., "Ancient DNA from South-East Europe Reveals Different Events during Early and Middle Neolithic Influencing the European Genetic Heritage," *PLoS One,* vol. 10, no. e0128810, 2015.

[259] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, et al., "Ancient human genomes suggest three ancestral populations for present-day Europeans," *Nature,* vol. 513, pp. 409-413, 2014.

[260] M. F. Deguilloux, L. Soler, M. H. Pemonge, C. Scarre, R. Joussaume and L. Laporte, "News From the West: Ancient DNA From a French Megalithic Burial Chamber.," *Am. J. Phys. Anthropol.,* vol. 144, p. 108–118, 2011.

[261] M. Lacan, C. Keyser, F.X. Ricaut, N. Brucato, J. Tarrús, et al., "Ancient DNA suggests the leading role played by men in the Neolithic dissemination," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 108, pp. 18255-18259, 2011.

Appendix A. Supporting Dataset to the paper in Chapter 5

Data S1

Neolithic mtDNA database with data grouped by regional cultures. Mean coordinates and dates of each culture are included at the end of its corresponding table. Cultures 1-8 and 11 are oldest local cultures with at least 9 individuals. They have been used to analyze the genetic cline of haplogroup K in Chapter 5, and are identified here with a different background color.

Ref.	Location	Country	Latitude	Longi-tude	cal BCE[5] mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	U	Middle PPNB	H28
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	HV	Middle PPNB	H53
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	K	Middle PPNB	H25
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	K	Middle PPNB	H4
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	K	Middle PPNB	H7
[165]	Tell Ramad	Syria	33.360	35.949	6975	7300	6650	K	PPNB	R65-14
[165]	Tell Ramad	Syria	33.360	35.949	6975	7300	6650	K	PPNB	R65-1S
[165]	Tell Ramad	Syria	33.360	35.949	6975	7300	6650	K	PPNB	R65-C8-SEB
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	H	Middle PPNB	H49
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	H	Middle PPNB	H68
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	R0	Middle PPNB	H3
[165]	Tell Ramad	Syria	33.360	35.949	6975	7300	6650	R0	PPNB	R64-4II
[165]	Tell Ramad	Syria	33.360	35.949	6975	7300	6650	R0	PPNB	R69
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	L3	Middle PPNB	H8
[165]	Tell Halula	Syria	36.417	38.167	7400	7500	7300	N	Middle PPNB	H70
1 Syria PPNB			**35.398**	**37.427**	**7258.3**	**7433.3**	**7083.3**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD

[5] cal BCE is the abbreviaton for calibrated Before Common Era. Common or Current Era (CE) is a year-numbering system that refers to the years since the start of this era (since AD 1). The preceding era is referred to as Before the Common Era (BCE), which is the one used here (and, therefore, in Chapter 5) because the Neolithic spread across Europe began 9,000 years ago.

[171]	Barcın	Turkey	40.300	29.567	6328.5	6419	6238	X2m	Anatolia Neolithic	Bar31
[171]	Barcın	Turkey	40.300	29.567	6121	6212	6030	K1a2	Anatolia Early Neolithic	Bar8
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	U8b1b1	Anatolia Early Neolithic	I0745
[80]	Barcın	Turkey	40.300	29.567	6350	6500	6200	H5	Anatolia Neolithic	I1580
[80]	Menteşe	Turkey	40.260	29.650	6000	6400	5600	N1a1a1	Anatolia Neolithic	I0725
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	N1a1a1	Anatolia Neolithic	I1096
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	N1a1a1a	Anatolia Neolithic	I0736
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	N1a1a1a	Anatolia Neolithic	I0854
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	T2b	Anatolia Neolithic	I1099
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	T2b	Anatolia Neolithic	I1101
[80]	Menteşe	Turkey	40.260	29.650	6000	6400	5600	H	Anatolia Neolithic	I0726
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	J1	Anatolia Neolithic	I1585
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	J1c11	Anatolia Neolithic	I0744
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	K1a or K1a1	Anatolia Neolithic	I0746
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	K1a or K1a6	Anatolia Neolithic	I1100
[80]	Menteşe	Turkey	40.260	29.650	6000	6400	5600	K1a2	Anatolia Neolithic	I0727
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	K1a2	Anatolia Neolithic	I1583
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	K1a3a	Anatolia Neolithic	I1102
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	K1a4	Anatolia Neolithic	I0707
[80]	Menteşe	Turkey	40.260	29.650	6000	6400	5600	K1a4	Anatolia Neolithic	I0724
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	K1a-C150T	Anatolia Neolithic	I1579
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	K1b1b1	Anatolia Neolithic	I1103
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	N1b1a	Anatolia Neolithic	I0708
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	U3	Anatolia Neolithic	I0709
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	U3	Anatolia Neolithic	I1581
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	W1-T119C	Anatolia Neolithic	I1097
[80]	Barcın	Turkey	40.300	29.567	6300	6400	6200	X2d2	Anatolia Neolithic	I1098
[80]	Menteşe	Turkey	40.260	29.650	6000	6400	5600	X2m2	Anatolia Neolithic	I0723

2 Anatolia	**40.293**	**29.582**	**6242.8**	**6397.5**	**6088.1**

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[172]	Vinkovci Nama	Croatia	45.286	18.798	5700	6000	5400	HV0	Starcevo	VINK3
[172]	Vinkovci Jugobanka	Croatia	45.283	18.794	5700	6000	5400	K	Starcevo	VINJ2
[172]	Vinkovci Jugobanka	Croatia	45.283	18.794	5700	6000	5400	T2b	Starcevo	VINJ1
[172]	Vukovar Gimnazija	Croatia	45.348	19.000	5700	6000	5400	T2b	Starcevo	VUKG4
[172]	Vinkovci Jugobanka	Croatia	45.283	18.794	5700	6000	5400	V	Starcevo	VINJ3
[172]	Vinkovci Nama	Croatia	45.286	18.798	5700	6000	5400	J1c	Starcevo	VINK2
[172]	Vukovar Gimnazija	Croatia	45.348	19.000	5700	6000	5400	J1c	Starcevo	VUKG1
[172]	Vukovar Gimnazija	Croatia	45.348	19.000	5700	6000	5400	J1c	Starcevo	VUKG3
[172]	Vinkovci Nama	Croatia	45.286	18.798	5700	6000	5400	K1a	Starcevo	VINK1
[172]	Vinkovci Nama	Croatia	45.286	18.798	5700	6000	5400	K1a	Starcevo	VINK5
[172]	Vinkovci Jugobanka	Croatia	45.283	18.794	5700	6000	5400	V6	Starcevo	VINJ4
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	H	Starcevo	BAM 10
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5590	5640	5540	H	Starcevo	BAM 11
[172]	Lánycsók, Gata-Csotola	Hungary	45.993	18.581	5700	6000	5400	H5	Starcevo	LGCS3
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	J	Starcevo	BAM 18
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5685	5740	5630	J1c	Starcevo	BAM 14
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5735	5810	5660	K	Starcevo	BAM 02
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5595	5650	5540	K	Starcevo	BAM 04
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	K	Starcevo	BAM 16
[172]	Lánycsók, Csata-alja	Hungary	45.996	18.581	5700	6000	5400	K	Starcevo	M6-116.4
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	K1	Starcevo	BAM 07
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	K1	Starcevo	BAM 24
[172]	Lánycsók, Gata-Csotola	Hungary	45.993	18.581	5700	6000	5400	K1a	Starcevo	LGCS4
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	K1a	Starcevo	BAM 09

[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	K1a	Starcevo	BAM 19
[172]	Lánycsók, Gata-Csotola	Hungary	45.993	18.581	5700	6000	5400	N1a1	Starcevo	LGCS2
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5630	5710	5550	N1a1a	Starcevo	BAM 22
[83]	Alsonyek-Bataszek, Mernoki telep	Hungary	46.200	18.700	5630	5710	5550	N1a1a1	Starcevo	I0174
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5750	5840	5660	T1a	Starcevo	BAM 17
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5585	5640	5530	T2	Starcevo	BAM 08
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5580	5640	5520	T2b	Starcevo	BAM 01
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	T2b	Starcevo	BAM 20
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5610	5680	5540	T2b	Starcevo	BAM 21
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5750	5840	5660	T2b	Starcevo	BAM 26
[172]	Lánycsók, Csata-alja	Hungary	45.996	18.581	5620	5680	5560	T2c	Starcevo	M6-116.9
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5545	5620	5470	T2e	Starcevo	BAM 05
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	U3	Starcevo	BAM 12
[172]	Lánycsók, Csata-alja	Hungary	45.996	18.581	5700	6000	5400	U4	Starcevo	M6-116.1
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5585	5650	5520	V	Starcevo	BAM 06
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	W	Starcevo	BAM 03
[172]	Lánycsók, Gata-Csotola	Hungary	45.993	18.581	5700	6000	5400	W	Starcevo	LGCS1
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5630	5710	5550	X2	Starcevo	BAM 13
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5700	6000	5400	X2	Starcevo	BAM 15
[172]	Alsónyék-Bátaszék, Mérnöki telep	Hungary	46.205	18.705	5560	5630	5490	X2	Starcevo	BAM 23
3 Hungary-Croatia Starcevo			**45.946**	**18.722**	**5674.5**	**5890.7**	**5458.4**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	H	LBK?	KAR 11
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	H	LBK?	KAR 20
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	H	LBK	KAR 29
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	H	LBK	KAR 59
[78]	Karsdorf	Germany	51.273	11.656	5049.5	5079	5020	H	LBK	KAR 18

[78, 76]	Karsdorf	Germany	51.273	11.656	5138.5	5207	5070	H1bz	LBK	KAR 6
[78, 80]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	H46b	LBK	KAR 16
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	HV	LBK?	KAR 17
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	J	LBK?	KAR 1
[78]	Naumburg	Germany	51.150	11.817	5137.5	5500	4775	J	LBK	NAU 2
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	J1c	LBK?	KAR 57
[78]	Karsdorf	Germany	51.273	11.656	5007.5	5056	4959	J1c	LBK	KAR 14
[78]	Karsdorf	Germany	51.273	11.656	5114.5	5140	5089	J1c2	LBK	KAR 3
[78]	Karsdorf	Germany	51.273	11.656	5030	5068	4992	K	LBK	KAR 10
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	K1a	LBK?	KAR 54
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	K1a	LBK?	KAR 7
[78]	Naumburg	Germany	51.150	11.817	5137.5	5500	4775	K1a	LBK	NAU 3
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	K1b1a	LBK?	KAR 8
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	K2a5	LBK?	KAR 55
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	N1a1a3	LBK?	KAR 40
[78]	Naumburg	Germany	51.150	11.817	5137.5	5500	4775	T2b	LBK	NAU 1
[78]	Karsdorf	Germany	51.273	11.656	5036	5075	4997	T2b	LBK	KAR 15
[78]	Naumburg	Germany	51.150	11.817	5137.5	5500	4775	T2c	LBK	NAU 5
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	T2e	LBK?	KAR 13
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	T2f	LBK?	KAR 9
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	U5a	LBK	KAR 4
[78]	Karsdorf	Germany	51.273	11.656	5137.5	5500	4775	U5b	LBK?	KAR 19
[78]	Oberwiederstedt 1, Unterwiederstedt	Germany	51.660	11.530	5137.5	5500	4775	T2f	LBK	UWS 11
[78]	Oberwiederstedt 1, Unterwiederstedt	Germany	51.660	11.530	5137.5	5500	4775	J	LBK	UWS 4
[76]	Unterwiederstedt	Germany	51.660	11.530	5139.5	5209	5070	J1c17	LBK	I0054
[78]	Oberwiederstedt 1, Unterwiederstedt	Germany	51.660	11.530	5137.5	5500	4775	J	LBK	UWS 8b
[78]	Oberwiederstedt 1, Unterwiederstedt	Germany	51.660	11.530	5137.5	5500	4775	N1a1a3	LBK	UWS 6
[78]	Oberwiederstedt 1, Unterwiederstedt	Germany	51.660	11.530	5137.5	5500	4775	T2b23a	LBK	UWS 7

[78]	Oberwiederstedt 1, Unterwiederstedt	Germany	51.660	11.530	5137.5	5500	4775	N1a1a3	LBK	UWS 5.2
[78, 76]	Oberwiederstedt 1, Unterwiederstedt	Germany	51.660	11.530	5137.5	5500	4775	K	LBK	UWS 3
[78, 76]	Oberwiederstedt 1, Unterwiederstedt	Germany	51.660	11.530	5137.5	5500	4775	K	LBK	UWS 2
4 Eastern Germany LBK			**51.356**	**11.642**	**5125.1**	**5425.9**	**4824.2**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5183	5207	5159	H	LBK	deb21
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	H	LBK	deb09
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5263	5300	5226	HV	LBK	deb20
[76]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5122	5171	5073	HV	LBK	deb05
[76]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5112	5185	5039	HV	LBK	deb04
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	J	LBK	deb26
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	J	LBK	deb30
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	J	LBK	deb37l
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	K	LBK	deb38
[76]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5075	5500	4650	K	LBK	deb02
[169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	4982	5020	4944	K	LBK	deb29ll
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	K1a	LBK	deb10
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	4978	5023	4933	N1a	LBK	deb22
[76]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5024	5064	4984	N1a1	LBK	deb01
[76]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5117	5186	5048	N1a1a	LBK	deb03
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	T	LBK	deb35ll
[169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5112	5185	5039	T	LBK	deb32
[169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	4951.5	4997	4906	T	LBK	deb11
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5183	5207	5159	T2	LBK	deb39
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	T2	LBK	deb15

[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	T2	LBK	deb33
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	U5a1a	LBK	deb36
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5039.5	5075	5004	V	LBK	deb12l
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	W	LBK	deb23
[78, 169]	Derenburg-Meerenstieg II	Germany	51.871	10.908	5137.5	5500	4775	W	LBK	deb34ll
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	H	LBK	HAL 11
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5053	5080	5026	H1e	LBK	HAL 39
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	H23	LBK	HAL 36
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	H26	LBK	HAL 32
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	J	LBK	HAL 35
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	K	LBK	HAL 12
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	K	LBK	HAL 18
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5122	5171	5073	K	LBK	HAL 31
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	K1a	LBK	HAL 20
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	K1a	LBK	HAL 9
[78, 83]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5129	5206	5052	K1a	LBK	I0048
[78, 80]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	K1a2	LBK	I1550
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5041	5079	5003	N1a1	LBK	HAL 7
[78, 83]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137	5207	5067	N1a1a1	LBK	I0057
[78, 83]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	4989	5032	4946	N1a1a1a	LBK	I0100
[78, 83]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5038	5079	4997	N1a1a1a2	LBK	I0659
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	N1a1a3	LBK	HAL 27
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	4989	5030	4948	N1a1a3	LBK	HAL 15
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	T2b	LBK	HAL 21
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	T2b	LBK	HAL 22
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	T2b	LBK	HAL 30
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	T2b	LBK	HAL 40

[78, 83]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5129	5206	5052	T2b	LBK	I0056
[78, 76]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5122	5171	5073	T2b	LBK	HAL 3
[78, 83]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5105	5206	5004	T2c1	LBK	I0046
[78, 76]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5183	5207	5159	V	LBK	HAL 1
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	V	LBK	HAL 16
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	V	LBK	HAL 17
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5137.5	5500	4775	V	LBK	HAL 38
[78]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	5272.5	5298	5247	W	LBK	HAL 37
[78, 83]	Halberstadt-Sonntagsfeld	Germany	51.890	11.040	4988	5034	4942	X2d1	LBK_EN	I0821
5 Western Germany LBK			**51.881**	**10.981**	**5114.8**	**5332.6**	**4896.9**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[166]	Can Sadurni	Spain	41.334	1.922	5390	5475	5305	H	Cardial Culture	CSA16
[166]	Can Sadurni	Spain	41.334	1.922	5390	5475	5305	K	Cardial Culture	CSA15 2223
[173]	Cova Bonica, Barcelona	Spain	41.370	1.894	5415	5470	5360	K1a2a	Early Neolithic	CB13
[166]	Can Sadurni	Spain	41.334	1.922	5390	5475	5305	N	Cardial Culture	CSA05 11
[166]	Can Sadurni	Spain	41.334	1.922	5390	5475	5305	N	Cardial Culture	CSA29
[166]	Can Sadurni	Spain	41.334	1.922	5390	5475	5305	X1	Cardial Culture	CSA26
[173]	Cova Bonica, Barcelona	Spain	41.370	1.894	5415	5470	5360	X2c	Early Neolithic	CB14
[166]	Cueva de Chaves	Spain	42.212	-0.138	5164	5329	4999	K	Cardial Culture	1CH01 02
[166]	Cueva de Chaves	Spain	42.212	-0.138	5164	5329	4999	H	Cardial Culture	2CH01 02
[166]	Cueva de Chaves	Spain	42.212	-0.138	5164	5329	4999	H	Cardial Culture	3CH01
[83]	Els Trocs	Spain	42.452	0.564	5258	5310	5206	N1a1a1	Iberia EN	Troc5
[83]	Els Trocs	Spain	42.452	0.564	5122	5178	5066	T2c1d	Iberia EN	Troc3
[83]	Els Trocs	Spain	42.452	0.564	5253.5	5303	5204	V	Iberia EN	Troc7
[83]	Els Trocs	Spain	42.452	0.564	5264.5	5311	5218	J1c3	Iberia EN	Troc1
[83]	Els Trocs	Spain	42.452	0.564	5122.5	5177	5068	K1a2a	Iberia EN	Troc4

6 North-Eastern Spain Cardial	**41.887**	**1.054**	**5286.2**	**5372.1**	**5200.3**	

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-173
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-182
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-193S
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-194
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-196
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-21
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-222
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-33
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-341
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-48
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-497
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	H	Early Neolithic farming	CAS-90
[78, 167]	Paternanbidea (Navarra)	Spain	42.795	-1.758	4967.5	5207	4728	H	Early Neolithic farming	PAT-1E3
[78, 167]	Paternanbidea (Navarra)	Spain	42.795	-1.758	4967.5	5207	4728	H	Early Neolithic farming	PAT-1E5
[78, 167]	Paternanbidea (Navarra)	Spain	42.795	-1.758	4967.5	5207	4728	H	Early Neolithic farming	PAT-2E1
[78, 167]	Paternanbidea (Navarra)	Spain	42.795	-1.758	4967.5	5207	4728	H3	Early Neolithic farming	PAT-1E4
[78, 167]	Paternanbidea (Navarra)	Spain	42.795	-1.758	4967.5	5207	4728	H3	Early Neolithic farming	PAT-4E2
[78, 167]	Paternanbidea (Navarra)	Spain	42.795	-1.758	4967.5	5207	4728	HV	Early Neolithic farming	PAT-3E2

[78, 167]	Paternanbidea (Navarra)	Spain	42.795	-1.758	4967.5	5207	4728	I	Early Neolithic farming	PAT-4E1
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	J	Early Neolithic farming	CAS-179
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	J	Early Neolithic farming	CAS-203
[78, 167]	Paternanbidea (Navarra)	Spain	42.795	-1.758	4967.5	5207	4728	K	Early Neolithic farming	PAT-2E2
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	K1a	Early Neolithic farming	CAS-181
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	K1a	Early Neolithic farming	CAS-202
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	K1a	Early Neolithic farming	CAS-191
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	T2	Early Neolithic farming	CAS-180
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	U	Early Neolithic farming	CAS-148
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	U	Early Neolithic farming	CAS-183
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	U	Early Neolithic farming	CAS-216
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	U	Early Neolithic farming	CAS-254
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	U	Early Neolithic farming	CAS-258
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	U	Early Neolithic farming	CAS-517
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	U	Early Neolithic farming	CAS-70
[78, 167]	Paternanbidea (Navarra)	Spain	42.795	-1.758	4967.5	5207	4728	U	Early Neolithic farming	PAT-1E1
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	U5	Early Neolithic farming	CAS-204
[78, 167]	Los Cascajos (Navarra)	Spain	42.559	-2.188	4932.5	5310	4555	X	Early Neolithic farming	CAS-257
7 Spain Navarre			**42.618**	**-2.081**	**4941.3**	**5284.3**	**4598.3**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD

[78, 174, 175]	Gruta do Caldeirão	Portugal	39.651	-8.414	5161.5	5480	4843	U	Neolithic Portugal	CALO1140
[78, 174, 175]	Gruta do Caldeirão	Portugal	39.651	-8.414	5161.5	5480	4843	V	Neolithic Portggal	CALP12130
[173]	Almonda cave	Portugal	39.505	-8.615	5265	5310	5220	H4a1a	Early Neolithic	F19
[78, 174, 175]	Gruta do Caldeirão	Portugal	39.651	-8.414	5161.5	5480	4843	H	Neolithic Portugal	CALN1424
[78, 174, 175]	Gruta do Caldeirão	Portugal	39.651	-8.414	5161.5	5480	4843	H	Neolithic Portugal	CALO1174
[78, 174, 175]	Gruta do Caldeirão	Portugal	39.651	-8.414	5161.5	5480	4843	H	Neolithic Portugal	CALO1436
[78, 174, 175]	Gruta do Caldeirão	Portugal	39.651	-8.414	5161.5	5480	4843	H	Neolithic Portugal	CALP11317
[78, 174, 175]	Gruta do Caldeirão	Portugal	39.651	-8.414	5161.5	5480	4843	H	Neolithic Portugal	CALP12151
[78, 174, 175]	Gruta do Caldeirão	Portugal	39.651	-8.414	5161.5	5480	4843	H	Neolithic Portugal	CALQ12181
[173]	Almonda cave	Portugal	39.505	-8.615	5280	5330	5230	H3	Early Neolithic	G21
8 Portugal coastal Early Neolithic			**39.621**	**-8.454**	**5183.7**	**5448.0**	**4919.4**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[258]	Cârcea	Romania	44.272	23.898	6000	6500	5500	H	Starcevo	Ca1
[258]	Cârcea	Romania	44.272	23.898	6000	6500	5500	T1a	Starcevo	Ca2
[258]	Gura Baciului	Romania	46.775	23.503	6000	6500	5500	J	Starcevo	GB2
[258]	Gura Baciului	Romania	46.775	23.503	6000	6500	5500	HV	Starcevo	GB3
[258]	Negrileşti	Romania	45.936	26.704	6000	6500	5500	H	Starcevo	NE-1
9 Romania Starcevo			**45.606**	**24.301**	**6000.0**	**6500.0**	**5500.0**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[259]	Viesenhaeuser Hof, Stuttgart-Muehlhausen	Germany	48.780	9.180	4950	5100	4800	T2c1d1	LBK	Stuttgart

[83]	Viesenhaeuser Hof, Stuttgart-Muehlhausen	Germany	48.780	9.180	5150	5500	4800	T2e	LBK	I0022
[83]	Viesenhaeuser Hof, Stuttgart-Muehlhausen	Germany	48.780	9.180	5150	5500	4800	T2b	LBK	I0025
[83]	Viesenhaeuser Hof, Stuttgart-Muehlhausen	Germany	48.780	9.180	5150	5500	4800	T2b	LBK	I0026
10 Southern Germany LBK			**48.780**	**9.180**	**5100.0**	**5400.0**	**4800.0**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[176]	Resmo	Sweden	56.538	16.446	2326	2451	2201	H	TRB	Res20
[176]	Linköping	Sweden	58.408	15.625	2501	2851	2151	N1a1a1a	BAC	Ber2
[176]	Resmo	Sweden	56.538	16.446	2651	2851	2451	J1d5	TRB	Res15
[176, 177]	Gökhem	Sweden	58.183	13.400	2751	2851	2651	H24	TRB	Gök7
[78, 176]	Gökhem	Sweden	58.183	13.400	3150	3400	2900	J2b1a	TRB	Ste9
[176, 177]	Gökhem	Sweden	58.183	13.400	2951	3101	2801	K1a5	TRB	Gök5
[78, 176, 177]	Gökhem	Sweden	58.183	13.400	3000	3100	2900	H	TRB	Gök4
[78, 176]	Gökhem	Sweden	58.183	13.400	3150	3400	2900	T2b	TRB	Ste7
[176, 177]	Gökhem	Sweden	58.183	13.400	2951	3101	2801	H	TRB	Gök2
11 Sweden			**57.842**	**14.324**	**2825.7**	**3011.8**	**2639.6**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	U	Middle Neolithic	Su13
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	U	Middle Neolithic	Su3
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	U4	Middle Neolithic	Su1
[258]	Curăteşti	Romania	44.277	26.830	4900	5300	4500	U5	Middle Neolithic	Cu1
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	U5b	Middle Neolithic	Su8
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	H5	Middle Neolithic	Va12
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	HV	Middle Neolithic	Va10
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	J	Middle Neolithic	Va2
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	J	Middle Neolithic	Va5

[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	J	Middle Neolithic	Va9
[258]	Curăteşti	Romania	44.277	26.830	4900	5300	4500	K	Middle Neolithic	Cu2
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	K	Middle Neolithic	Su4
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	H	Middle Neolithic	BV1
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	H	Middle Neolithic	BV2
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	H	Middle Neolithic	Su11
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	H	Middle Neolithic	Su12
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	H	Middle Neolithic	Su14
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	H	Middle Neolithic	Su15
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	H	Middle Neolithic	Su16
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	H	Middle Neolithic	Su7
[258]	Sultana-Valea-Orbului	Romania	44.259	26.853	4900	5300	4500	H	Middle Neolithic	Su9
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	H	Middle Neolithic	Va11
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	H	Middle Neolithic	Va3
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	H	Middle Neolithic	Va4
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	H	Middle Neolithic	Va6
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	H	Middle Neolithic	Va8
[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	R	Middle Neolithic	SMR-2
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	T1a	Middle Neolithic	Va1
[258]	Vărăşti	Romania	44.237	26.248	5000	5500	4500	W6	Middle Neolithic	Va7
12 Romania Middle Neolithic			**44.250**	**26.559**	**4925.9**	**5369.0**	**4482.8**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	H	Middle Neolithic	SMR-1
[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	H	Middle Neolithic	SMR-10
[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	H	Middle Neolithic	SMR-3
[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	H	Middle Neolithic	SMR-4
[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	H	Middle Neolithic	SMR-5
[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	H	Middle Neolithic	SMR-6

[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	H	Middle Neolithic	SMR-7
[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	H	Middle Neolithic	SMR-8
[258]	Sultana-Malu Roşu	Romania	44.259	26.853	4250	4500	4000	H	Middle Neolithic	SMR-9
13 Romania Late-Middle Neolithic			**44.259**	**26.853**	**4250.0**	**4500.0**	**4000.0**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5200	5500	4900	H	LBK	BUD 10
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5200	5500	4900	H	LBK	BUD 7
[172]	M85 Enese elkerülő 02. Kóny, Proletár-dűlő II	Hungary	47.639	17.365	5200	5500	4900	H	LBK	KON 1
[80, 75]	Debrecen Tocopart Erdoalja	Hungary	47.520	21.589	5217.5	5291	5144	H	Alföld Linear Pottery	I1498
[172]	Balatonszemes-Bagódomb	Hungary	46.789	17.785	5200	5500	4900	H	LBK	BAB 3
[172]	Balatonszemes-Bagódomb	Hungary	46.789	17.785	5200	5500	4900	H	LBK	BAB 5
[172]	Balatonszemes-Bagódomb	Hungary	46.789	17.785	5200	5500	4900	H	LBK	BAB 6
[172]	Bölcske-Gyűrűsvölgy	Hungary	46.767	18.879	5200	5500	4900	H	LBK	BÖVÖ 1
[172]	Bölcske-Gyűrűsvölgy	Hungary	46.767	18.879	5200	5500	4900	H	LBK	BÖVÖ 3
[172]	Balatonszárszó-Kis-erdei-dűlő	Hungary	46.820	17.858	5200	5500	4900	H	LBK	BSZ 15
[172]	Balatonszemes-Bagódomb	Hungary	46.789	17.785	5200	5500	4900	H26b	LBK	BAB 4
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	4955	5060	4850	H5	LBK	BUD 13
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5200	5500	4900	H5	LBK	BUD 15
[172]	Balatonszárszó-Kis-erdei-dűlő	Hungary	46.820	17.858	5200	5500	4900	HV	LBK	BSZ 19
[172]	Bölcske-Gyűrűsvölgy	Hungary	46.767	18.879	5200	5500	4900	J	LBK	BÖVÖ 2
[172]	Balatonszárszó-Kis-erdei-dűlő	Hungary	46.820	17.858	5200	5500	4900	J	LBK	BSZ 21
[172]	Balatonszárszó-Kis-erdei-dűlő	Hungary	46.820	17.858	5200	5500	4900	J	LBK	BSZ 9
[80, 75]	Kompolt-Kigyoser	Hungary	47.167	20.833	5205.5	5295	5116	J1c1	Late Alföld Linear Pottery	I1500

[80, 75]	Polgar Ferenci hat	Hungary	47.880	21.192	5163.5	5211	5116	J1c5	Tiszadob-Bükk Culture	I1505
[172]	Bölcske-Gyűrűsvölgy	Hungary	46.767	18.879	5200	5500	4900	K	LBK	BÖVÖ 4
[172]	Harta-Gátőrház	Hungary	46.705	19.015	5200	5500	4900	K	LBK	HARG 4
[172]	Tolna-Mözs	Hungary	46.407	18.742	5190	5310	5070	K	LBK	TOLM 4
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5115	5220	5010	K1a	LBK	BUD 14
[172]	Bölcske-Gyűrűsvölgy	Hungary	46.767	18.879	5200	5500	4900	K1a	LBK	BÖVÖ 5
[80, 75]	Apc-Berekalya I	Hungary	47.167	19.833	5104	5206	5002	K1a3a3	LBK	I1496
[172]	Harta-Gátőrház	Hungary	46.705	19.015	5200	5500	4900	N1a1a	LBK	HARG 2
[83]	Szemely-Hegyes	Hungary	46.400	18.740	5075	5210	4940	N1a1a1a3	Hungary EN	I0176
[172]	Balatonszárszó-Kis-erdei-dűlő	Hungary	46.820	17.858	5200	5500	4900	N1a1a3	LBK	BSZ 5
[172]	Szemely-Hegyes	Hungary	46.026	18.323	5105	5210	5000	N1a1a3	LBK	SZEH 9
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5200	5500	4900	T1a	LBK	BUD 4
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5200	5500	4900	T2	LBK	BUD 5
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5200	5500	4900	T2b	LBK	BUD 12
[172]	M85 Enese elkerülő 02. Kóny, Proletár-dűlő II	Hungary	47.639	17.365	5200	5500	4900	T2b	LBK	KON 3
[172]	Harta-Gátőrház	Hungary	46.705	19.015	5200	5500	4900	T2b	LBK	HARG 1
[172]	Harta-Gátőrház	Hungary	46.705	19.015	5200	5500	4900	T2b	LBK	HARG 5
[172]	M85 Enese elkerülő 02. Kóny, Proletár-dűlő II	Hungary	47.639	17.365	4940	5050	4830	T2b23a	LBK	KON 5
[172]	M85 Enese elkerülő 02. Kóny, Proletár-dűlő II	Hungary	47.639	17.365	5200	5500	4900	T2b23a	LBK	KON 4
[172]	Harta-Gátőrház	Hungary	46.705	19.015	5200	5500	4900	T2c	LBK	HARG 3
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5130	5220	5040	T2e	LBK	BUD 3
[172]	Tolna-Mözs	Hungary	46.407	18.742	5105	5210	5000	T2e	LBK	TOLM 3
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5200	5500	4900	U2	LBK	BUD 9

[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5200	5500	4900	U5a1	LBK	BUD 1
[80, 75]	Polgar Ferenci hat	Hungary	47.880	21.192	5267.5	5306	5229	U5b2c	Alföld Linear Pottery	I1506
[172]	Budakeszi 4/8 Szőlőskert-Tangazdaság	Hungary	47.502	18.910	5130	5220	5040	V	LBK	BUD 2
[80, 75]	Garadna	Hungary	48.520	21.168	5206.5	5281	5132	X2b-T226C	Bükk Culture	I1499
14 Hungary LBK			**47.124**	**18.814**	**5175.8**	**5406.7**	**4944.9**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78]	Esperstedt	Germany	51.420	11.680	4628.5	4705	4552	T2e	RSC	ESP 13
[78]	Oberwiederstedt 3, Schrammhoehe	Germany	51.660	11.530	4619	4686	4552	T2f	RSC	OSH 8
[78]	Oberwiederstedt 3, Schrammhoehe	Germany	51.660	11.530	4494.5	4582	4407	H5b	RSC	OSH 7
[78]	Oberwiederstedt 3, Schrammhoehe	Germany	51.660	11.530	4437.5	4625	4250	HV0	RSC	OSH 10
[78]	Oberwiederstedt 3, Schrammhoehe	Germany	51.660	11.530	4437.5	4625	4250	K	RSC	OSH 6
[78]	Oberwiederstedt 4, Arschkerbe Ost	Germany	51.660	11.530	4437.5	4625	4250	N1a1a	RSC	OAO 1
[78]	Oberwiederstedt 3, Schrammhoehe	Germany	51.660	11.530	4437.5	4625	4250	H1	RSC	OSH 3
[78]	Oberwiederstedt 3, Schrammhoehe	Germany	51.660	11.530	4437.5	4625	4250	H16a	RSC	OSH 1
[78]	Oberwiederstedt 3, Schrammhoehe	Germany	51.660	11.530	4437.5	4625	4250	H89	RSC	OSH 2
[78]	Oberwiederstedt 3, Schrammhoehe	Germany	51.660	11.530	4437.5	4625	4250	X2c	RSC	OSH 5
15 Eastern Germany RSC			**51.636**	**11.545**	**4480.5**	**4634.8**	**4326.1**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	H	SCG	SALZ 28
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4006	4045	3967	H	SCG	SALZ 38
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3983.5	4004	3963	H	SCG	SALZ 107
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3675	3950	3400	H	BAC	SALZ 55
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4130.5	4172	4089	H10	SCG	SALZ 18
[83]	Esperstedt	Germany	51.422	11.676	3842	3887	3797	H1e1a	BAC	ESP 30
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	H1e7	SCG	SALZ 21

[78]	Halle-Queis	Germany	51.480	12.130	3675	3950	3400	H7d5	BAC	HQU 4
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4009.5	4034	3985	HV	SCG	SALZ 24
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	J	SCG	SALZ 10
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	J1c	SCG	SALZ 11
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	J1c	SCG	SALZ 42
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3983.5	4004	3963	J1c	SCG	SALZ 110
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	J2b1a	SCG	SALZ 12
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4148.5	4171	4126	K	SCG	SALZ 30
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4148.5	4171	4126	K	SCG	SALZ 31
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	K	SCG	SALZ 40
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	K	SCG	SALZ 9
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	K1a	SCG	SALZ 13
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	K1a	SCG	SALZ 14
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	K1a	SCG	SALZ 15
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	K1a	SCG	SALZ 22
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	K1a	SCG	SALZ 41
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	K1a	SCG	SALZ 8
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	N1a1a3	SCG	SALZ 25
[78]	Karsdorf	Germany	51.273	11.656	3675	3950	3400	T1a1'3	BAC	KAR 22
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	T2b	SCG	SALZ 43
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	T2b	SCG	SALZ 44
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	T2c	SCG	SALZ 32
[78]	Karsdorf	Germany	51.273	11.656	3675	3950	3400	T2c	BAC	KAR 21
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	T2f	SCG	SALZ 34
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	U5b2a2c	SCG	SALZ 29
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3982	3996	3968	U5b3	SCG	SALZ 27
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	U8b1b	SCG	SALZ 39
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4095	4134	4056	W1c	SCG	SALZ 19
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	W1c	SCG	SALZ 20
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	W1c	SCG	SALZ 35

[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	4025	4100	3950	X2b1'2'3'4'5'6	SCG	SALZ 26
16 Eastern Germany SCG/BAC			**51.503**	**11.844**	**3990.2**	**4074.2**	**3906.3**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	H	SMC	SALZ 116
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	H	SMC	SALZ 66
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3298	3334	3262	H3	SMC	SALZ 57
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	H3	SMC	SALZ 77
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	H5	SMC	SALZ 1
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	H5	SMC	SALZ 118
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	H5	SMC	SALZ 5
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	H5	SMC	SALZ 6
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	H5	SMC	SALZ 7
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	HV	SMC	SALZ 48
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	J	SMC	SALZ 54
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	J1c	SMC	SALZ 52
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	J1c	SMC	SALZ 74
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	J1c	SMC	SALZ 84
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3204	3237	3171	J1c	SMC	SALZ 88
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	J2b1a	SMC	SALZ 78
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	K1	SMC	SALZ 49
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	K1a	SMC	SALZ 70
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3302	3339	3265	K1a4a1a2	SMC	SALZ 82
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	N1a1a3	SMC	SALZ 67
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	N1a1a3	SMC	SALZ 90
[83]	Esperstedt	Germany	51.422	11.676	3223	3360	3086	T2b	SMC	ESP 24
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	T2b	SMC	SALZ 63
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3273	3310	3236	U3a	SMC	SALZ 60

[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	U3a	SMC	SALZ 3
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	U3a	SMC	SALZ 4
[80]	Salzmuende-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	U3a1	SMC	I0551
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3274	3330	3218	U5b	SMC	SALZ 89
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	V	SMC	SALZ 2
[78]	Salzmünde-Schiebzig	Germany	51.520	11.852	3212.5	3400	3025	X2b1'2'3'4'5'6	SMC	SALZ 61
17 Eastern Germany SMC			**51.517**	**11.846**	**3222.5**	**3383.7**	**3061.3**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78]	Quedlinburg IX	Germany	51.790	11.140	3675	3950	3400	H	BAC	QLB 14
[78]	Quedlinburg VII 2	Germany	51.792	11.147	3675	3950	3400	H	BAC	QLB 4
[78]	Quedlinburg IX	Germany	51.790	11.140	3675	3950	3400	HV	BAC	QLB 15
[78]	Quedlinburg IX	Germany	51.790	11.140	3575	3640	3510	J	BAC	QLB 13
[78]	Quedlinburg VII 2	Germany	51.792	11.147	3675	3950	3400	K1a	BAC	QLB 1
[78]	Quedlinburg VII 2	Germany	51.792	11.147	3675	3950	3400	K1a	BAC	QLB 5
[78]	Quedlinburg VII 2	Germany	51.792	11.147	3660	3700	3620	N1a1a	BAC	QLB 8
[78]	Quedlinburg IX	Germany	51.790	11.140	3415	3460	3370	T2b	BAC	QLB 17
[78]	Quedlinburg VII 3	Germany	51.792	11.147	3675	3950	3400	T2c	BAC	QLB 9
[78]	Quedlinburg IX	Germany	51.790	11.140	3575	3640	3510	T2e	BAC	QLB 18
[78]	Quedlinburg VII 2	Germany	51.792	11.147	3675	3950	3400	U5b2a2	BAC	QLB 6
[78]	Quedlinburg VII 2	Germany	51.792	11.147	3675	3950	3400	U8a1a	BAC	QLB 2
[78]	Quedlinburg VII 2	Germany	51.792	11.147	3675	3950	3400	X	BAC	QLB 7
[78]	Quedlinburg IX	Germany	51.790	11.140	3670	3710	3630	X2c	BAC	QLB 11
18 Western Germany BAC			**51.791**	**11.144**	**3640.7**	**3835.7**	**3445.7**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	H	BEC	BENZ 17
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	H	BEC	BENZ 36
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	H1e1a3	BEC	BENZ 40
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	H5	BEC	BENZ 29
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	K1	BEC	BENZ 33
[78]	Benzingerode I	Germany	51.830	10.860	3010	3101	2919	K1a	BEC	BENZ 3
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	K1a	BEC	BENZ 27
[78]	Benzingerode I	Germany	51.830	10.860	3010	3101	2919	T2b	BEC	BENZ 6

[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	T2b	BEC	BENZ 19
[78]	Benzingerode I	Germany	51.830	10.860	3174.5	3251	3098	U5a	BEC	BENZ 20
[78]	Benzingerode I	Germany	51.830	10.860	3011.5	3104	2919	U5a	BEC	BENZ 14
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	U5b	BEC	BENZ 35
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	U5b1c1	BEC	BENZ 1
[78]	Benzingerode I	Germany	51.830	10.860	3010	3101	2919	U5b2a1a	BEC	BENZ 18
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	V	BEC	BENZ 39
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	W	BEC	BENZ 15
[78]	Benzingerode I	Germany	51.830	10.860	2875	3100	2650	X	BEC	BENZ 37
19 Western Germany BEC			**51.830**	**10.860**	**2924.5**	**3109.3**	**2739.6**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[260]	Prissé-la-Charrière	France	46.153	-0.484	4206	4336	4076	U5b	Neolithic	Prissé 2
[260]	Prissé-la-Charrière	France	46.153	-0.484	4257.5	4340	4175	N1a	Neolithic	Prissé 4
[260]	Prissé-la-Charrière	France	46.153	-0.484	4255.5	4340	4171	X2	Neolithic	Prissé 1
20 Western France Prissé			**46.153**	**-0.484**	**4239.7**	**4338.7**	**4140.7**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	H1	TRE	TRE 593
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	H1	TRE	TRE 596
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	H1	TRE	TRE 603
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	H3	TRE	TRE 577
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	H3	TRE	TRE 581
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	H3	TRE	TRE 600
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	HV0	TRE	TRE 573
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	HV0	TRE	TRE 609
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	J1	TRE	TRE 139
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	J1	TRE	TRE 209

[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	J1	TRE	TRE 583
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	J1	TRE	TRE 587
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	J1	TRE	TRE 612
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	J1	TRE	TRE 616
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	K1a	TRE	TRE 604
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	K1a	TRE	TRE 614
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	T2b	TRE	TRE 584
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	T2b	TRE	TRE 588
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	U	TRE	TRE 571
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	U5	TRE	TRE 137
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	U5	TRE	TRE 195
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	U5	TRE	TRE 575
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	U5	TRE	TRE 579
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	U5b1c	TRE	TRE 611
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	V	TRE	TRE 637
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	X2	TRE	TRE 570
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	X2	TRE	TRE 592
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	X2	TRE	TRE 615
[78, 253]	Treilles	France	43.930	3.027	2960	3030	2890	X2	TRE	TRE 636
21 South-Eastern France Treilles			**43.930**	**3.027**	**2960.0**	**3030.0**	**2890.0**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78, 261]	Avellaner cave	Spain	42.058	2.539	4760.7	5069.5	4452	H3	Epicardial Culture	AVE03
[78, 261]	Avellaner cave	Spain	42.058	2.539	4760.7	5069.5	4452	U5	Epicardial Culture	AVE07
[78, 261]	Avellaner cave	Spain	42.058	2.539	4760.7	5069.5	4452	T2b	Epicardial Culture	AVE04
[78, 261]	Avellaner cave	Spain	42.058	2.539	4760.7	5069.5	4452	T2b	Epicardial Culture	AVE05
[78, 261]	Avellaner cave	Spain	42.058	2.539	4760.7	5069.5	4452	K1a	Epicardial Culture	AVE01

[78, 261]	Avellaner cave	Spain	42.058	2.539	4760.7	5069.5	4452	K1a	Epicardial Culture	AVE02
[78, 261]	Avellaner cave	Spain	42.058	2.539	4760.7	5069.5	4452	K1a	Epicardial Culture	AVE06
22 Catalonia Epicardial			**42.058**	**2.539**	**4760.8**	**5069.5**	**4452.0**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[166]	Sant Pau del Camp	Spain	41.376	2.169	3975	4250	3700	K	Epicardial Culture	6SP 0102
[166]	Sant Pau del Camp	Spain	41.376	2.169	3975	4250	3700	H20	Epicardial Culture	26SP 0102
[166]	Sant Pau del Camp	Spain	41.376	2.169	3975	4250	3700	N	Epicardial Culture	27SP 0102
23 Catalonia Late Epicardial			**41.376**	**2.169**	**3975.0**	**4250.0**	**3700.0**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78, 167]	Marizulo (Gipuzkoa)	Spain	43.247	-1.991	4144	4315	3973	U5	Neolithic farming	MZ-1
[78, 167]	Fuente Hoz (Arava)	Spain	42.802	-2.898	4019	4330	3708	H	Neolithic farming	FH-3
[78, 167]	Fuente Hoz (Arava)	Spain	42.802	-2.898	4019	4330	3708	H	Neolithic farming	FH-6
[78, 167]	Fuente Hoz (Arava)	Spain	42.802	-2.898	4019	4330	3708	U5a	Neolithic farming	FH-2
[78, 167]	Fuente Hoz (Arava)	Spain	42.802	-2.898	4019	4330	3708	U	Neolithic farming	FH-1
[78, 167]	Fuente Hoz (Arava)	Spain	42.802	-2.898	4019	4330	3708	U	Neolithic farming	FH-4
[78, 167]	Fuente Hoz (Arava)	Spain	42.802	-2.898	4019	4330	3708	U	Neolithic farming	FH-5
24 Spain Basque country			**42.865**	**-2.768**	**4036.9**	**4327.9**	**3745.9**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78, 174, 175]	Algar do Bom Santo	Portugal	39.145	-9.019	3490	3630	3350	H	Neolithic Portugal	ABS.AE 1.521
[78, 174, 175]	Algar do Bom Santo	Portugal	39.145	-9.019	3490	3630	3350	H	Neolithic Portugal	ABS.BB 4.250
[78, 174, 175]	Algar do Bom Santo	Portugal	39.145	-9.019	3490	3630	3350	U5a1a	Neolithic Portugal	ABS.AE 2.175
25 Portugal coastal Late Neolithic			**39.145**	**-9.019**	**3490.0**	**3630.0**	**3350.0**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[78, 174, 175]	Perdigões	Portugal	38.430	-7.534	3250	3500	3000	U5a1a	Neolithic Portugal	Perd1
[78, 174, 175]	Perdigões	Portugal	38.430	-7.534	3250	3500	3000	U5a1a	Neolithic Portugal	Perd2
[78, 174, 175]	Perdigões	Portugal	38.430	-7.534	3250	3500	3000	H	Neolithic Portugal	Perd5
[78, 174, 175]	Perdigões	Portugal	38.430	-7.534	3250	3500	3000	H	Neolithic Portugal	Perd6
[78, 174, 175]	Perdigões	Portugal	38.430	-7.534	3250	3500	3000	H	Neolithic Portugal	Perd7
[78, 174, 175]	Perdigões	Portugal	38.430	-7.534	3250	3500	3000	H	Neolithic Portugal	Perd8
26 Portugal inland Late Neolithic			**38.430**	**-7.534**	**3250.0**	**3500.0**	**3000.0**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[173]	Cova de la Sarsa, València	Spain	38.760	-0.582	5274	5321	5227	K1a4a1	Early Neolithic	CS7675
[173]	Cova de l'Or, Alicante	Spain	38.845	-0.364	5335	5360	5310	H4a1a	Early Neolithic	H3C6
Spain, Valencia (not taken into account)			**38.802**	**-0.473**	**5304.5**	**5340.5**	**5268.5**			

Ref.	Location	Country	Latitude	Longi-tude	cal BCE mean	cal BCE max	cal BCE min	mtDNA HG	Archeol context	sample iD
[171]	Paliambela	Greece	40.511	22.507	4401	4452	4350	J1c1	Late Neolithic	Pal7
[171]	Kleitos	Greece	40.433	21.854	4112.5	4230	3995	K1a2	Final Neolithic	Klei10
[171]	Revenia	Greece	39.488	20.918	6351	6438	6264	X2b	Early Neolithic	Rev5
Greece (not taken into account)			**40.144**	**21.760**	**4954.8**	**5040.0**	**4869.7**			

Data S2

Numbers of individuals with known mtDNA and classified by regional culture and haplogroup. These numbers have been computed from the tables in Data.

Region	Number of individuals													
	K	N1a	HV	R0	H5	T2	J	U3	W	X	V	other		total
1 Syria PPNB	6	0	1	3	0	0	0	0	0	0	0	5	U, L3,N, H	15
2 Anatolia	10	5	0	0	1	2	2	2	1	3	0	2	U8, H	28
3 Hungary-Croatia Starcevo	12	3	1	0	1	9	5	1	2	3	3	4	U4, T1a, H	44
4 Eastern Germany LBK	8	3	1	0	0	7	8	0	0	0	0	9	H, H1, U5	36
5 Western Germany LBK	11	9	3	0	0	10	4	0	3	1	5	10	H, H1, H2, T; U5	56
6 North-Eastern Spain Cardial	4	1	0	0	0	1	1	0	0	2	1	5	H, N	15
7 Spain Navarre	4	0	1	0	0	1	2	0	0	1	0	27	H, H3, U, U5	36
8 Portugal coastal Early Neolithic	0	0	0	0	0	0	0	0	0	0	1	9	H, H3, H4, U	10
9 Romania Starcevo	0	0	1	0	0	0	1	0	0	0	0	3	H, T1a	5
10 Southern Germany LBK	0	0	0	0	0	4	0	0	0	0	0	0	-	4
11 Sweden	1	1	0	0	0	1	2	0	0	0	0	4	H, H24	9
12 Romania Middle Neolithic	2	0	1	1	1	0	3	0	1	0	0	20	U, U4, U5, H, T1a	29
13 Romania Late-Middle Neolithic	0	0	0	0	0	0	0	0	0	0	0	9	H	9
14 Hungary LBK	6	4	1	0	2	10	5	0	0	1	1	15	H, T1a, U2, U5	45
15 Eastern Germany RSC	1	1	1	0	1	2	0	0	0	1	0	3	H, H16, H89	10
16 Eastern Germany SCG/BAC	10	1	1	0	0	5	5	0	3	1	0	12	H, H10, T1a, U5, U8	38
17 Eastern Germany SMC	3	2	1	0	5	2	6	4	0	1	1	5	H, H3, U5	30
18 Western Germany BAC	2	1	1	0	0	3	1	0	0	2	0	4	H, U5, U8	14
19 Western Germany BEC	3	0	0	0	1	2	0	0	1	1	1	8	H, H1, U5a, U5b	17
20 Western France Prissé	0	1	0	0	0	0	0	0	0	1	0	1	U5b	3
21 South-Eastern France Treilles	2	0	2	0	0	2	6	0	0	4	1	12	H1, H3, U, U5	29
22 Catalonia Epicardial	3	0	0	0	0	2	0	0	0	0	0	2	H3, U5	7
23 Catalonia Late Epicardial	1	0	0	0	0	0	0	0	0	0	0	2	H20, N	3
24 Spain Basque country	0	0	0	0	0	0	0	0	0	0	0	7	H, U,U5	7
25 Portugal coastal Late Neolithic	0	0	0	0	0	0	0	0	0	0	0	3	H, U5, U8	3

26 Portugal inland Late Neolithic	0	0	0	0	0	0	0	0	0	0	0	6	H, U5	6
Total	89	32	16	4	12	63	51	7	11	22	14	187		508

Data S3

%K from S2 Data for the 26 regional cultures. Mean dates and locations, from S1 Data, used in Fig. 5.1.

	Haplogroup K (%)	Number of individuals	Latitude	Longitude	Distance to Ras Shamra (km)	Time (years BCE)	Time error (± years)
1 Syria PPNB	40.0	15	35.3978	37.4274	155.01	7258.33	175.00
2 Anatolia	35.7	28	40.2923	29.5827	751.42	6244.23	154.70
3 Hungary-Croatia Starcevo	27.3	44	45.9459	18.7219	1831.91	5674.55	216.14
4 Eastern Germany LBK	22.2	36	51.3561	11.6423	2596.33	5125.08	300.86
5 Western Germany LBK	19.6	56	51.8814	10.9812	2666.47	5114.77	217.84
6 North-Eastern Spain Cardial	26.7	15	41.8868	1.0537	3066.24	5286.00	85.90
7 Spain Navarre	11.1	36	42.6178	-2.0807	3325.90	4941.25	343.00
8 Portugal coastal Early Neolithic	0.0	10	39.6214	-8.4542	3879.19	5183.70	264.30
9 Romania Starcevo	0.0	5	45.6062	24.3012	1471.46	6000.00	500.00
10 Southern Germany LBK	0.0	4	48.7800	9.1800	2613.04	5088.00	287.50
11 Sweden	11.1	9	57.8425	14.3240	2937.19	2825.67	186.11
12 Romania Middle Neolithic	6.9	29	44.2500	26.5592	1239.75	4925.86	443.10
13 Romania Late-Middle Neolithic	0.0	9	44.2594	26.8531	1225.01	4250.00	250.00
14 Hungary LBK	13.3	45	47.1238	18.8142	1900.66	5176.00	230.90
15 Eastern Germany RSC	10.0	10	51.6360	11.5450	2619.05	4480.45	154.35
16 Eastern Germany SCG/BAC	26.3	38	51.5034	11.8444	2593.66	3990.24	83.92
17 Eastern Germany SMC	10.0	30	51.5168	11.8461	2594.38	3222.47	161.20
18 Western Germany BAC	14.3	14	51.7909	11.1441	2651.59	3640.71	195.00
19 Western Germany BEC	17.6	17	51.8300	10.8600	2670.34	2924.47	184.82
20 Western France Prissé	0.0	3	46.1527	-0.4844	3230.44	4239.67	99.00
21 South-Eastern France Treilles	6.9	29	43.9303	3.0273	2923.48	2960.00	70.00
22 Catalonia Epicardial	42.9	7	42.0578	2.5389	2944.35	4760.75	308.75
23 Catalonia Late Epicardial	33.3	3	41.3761	2.1694	2972.00	3975.00	275.00
24 Spain Basque country	0.0	7	42.8652	-2.7683	3382.41	4036.86	291.00
25 Portugal coastal Late Neolithic	0.0	3	39.1447	-9.0186	3934.71	3490.00	140.00
26 Portugal inland Late Neolithic	0.0	6	38.4295	-7.5341	3819.67	3250.00	250.00

Data S4

Archaeological and simulated arrival times at 8 cultural regions for different sea-travel ranges. Values used in Fig. 5.12 (see Sec. 5.8.6).

Distance to Ras Shamra (km)	Time (mean cal BCE)	Iterations from oldest date (Ras Shamra)	Iterations from the simulations					Error between archaeological data and simulation (%)					
			No sea travel	max jump 50 km	max jump 100 km	max jump 150 km	max jump 200 km	No sea travel	max jump 50 km	max jump 100 km	max jump 150 km	max jump 200 km
0.00	8233	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00
313.99	7361	27	10	10	9	8	7	63.30	63.30	66.97	70.64	74.31
1770.21	6044	68	60	59	52	38	28	12.29	13.75	23.98	44.45	59.07
2456.79	5920	72	78	78	72	59	48	7.91	7.91	0.39	18.37	33.59
2673.40	5811	76	87	87	81	68	56	14.95	14.95	7.02	10.16	26.01
3125.24	5661	80	106	105	89	64	50	31.88	30.64	10.73	20.37	37.79
3265.32	5357	90	106	106	89	65	52	17.94	17.94	0.97	27.68	42.14
3934.86	5620	82	129	129	111	82	66	57.98	57.98	35.94	0.42	19.17
							Total error	206.25	206.47	146.00	192.09	292.09

Oldest Neolithic genetic regional data	Oldest archaeological site nearby	Latitude	Longitude	Nearest node of the grid
1 Syria PPNB	Ras Shamra	35.58	35.730	5512
2 Anatolia	Hayaz Höyük	37.48	38.330	6416
3 Hungary-Croatia Starcevo	Gudnja	42.917	17.417	9263
4 Eastern Germany LBK	Dresden-Prohlis	51.051	13.738	12322
5 Western Germany LBK	Eilsleben	52.167	11.167	13400
6 North-Eastern Spain Cardial	Forcas	42.180	0.351	10497
7 Spain Navarre	Aizpea (Basque Country)	42.943	-1.328	10856
8 Portugal coastal Early Neolithic	Vale Pincel I	37.946	-8.7672	10120

Data S5

Spread in homogeneous space. Values used in Fig. 5.24 (see Sec. 5.8.13).

η	front speed [km/year]		
	Eq. (S60)	simulation	error [%]
0	1,1047	1,1345	2,6935
0,01	1,1100	1,1388	2,5907
0,02	1,1153	1,1439	2,5697
0,05	1,1304	1,1595	2,5760
0,1	1,1539	1,1858	2,7599
0,2	1,1956	1,2309	2,9555
0,35	1,2473	1,2925	3,6200
0,5	1,2895	1,3435	4,1879
0,75	1,3445	1,4098	4,8548
1	1,3862	1,4605	5,3624

Data S6

Fraction of the population with haplogroup K in each regional culture (from S3 Data) and their 80% confidence-level (CL) ranges. Values on white background have at least 9 individuals, and they have been used in Figs. 5.2-5.3.

Oldest Neolithic genetic regional data (including Sweden)	%K	# individuals	CL 80% lower bound	CL 80% upper bound	CL 80% error -	CL 80% error +	Distance to Ras Shamra (km)
1 Syria PPNB	40.00	15	26.67	53.33	13.33	13.33	155.01
2 Anatolia	35.71	28	25.00	46.43	10.71	10.71	751.42
3 Hungary-Croatia Starcevo	27.27	44	18.18	36.36	9.09	9.09	1831.91
4 Eastern Germany LBK	22.22	36	13.89	30.56	8.33	8.33	2596.33
5 Western Germany LBK	19.64	56	12.50	26.79	7.14	7.14	2666.47
6 North-Eastern Spain Cardial	26.67	15	13.33	40.00	13.33	13.33	3066.24
7 Spain Navarre	11.11	36	5.56	16.67	5.56	5.56	3325.90
8 Portugal coastal Early Neolithic	0.00	10	0.00	14.00	0.00	14.00	3879.19
9 Romania Starcevo	0.00	5					1471.46
10 Southern Germany LBK	0.00	4					2613.04
11 Sweden	11.11	9	0.00	22.22	11.11	11.11	2937.19
More recent genetic regional data							
12 Romania Middle Neolithic	6.90	29	0.00	13.79	6.90	6.90	1239.75
13 Romania Late-Middle Neolithic	0.00	9	0.00	15.00	0.00	0.15	1225.01
14 Hungary LBK	13.33	45	6.67	20.00	6.67	6.67	1900.66
15 Eastern Germany RSC	10.00	10	0.00	20.00	10.00	10.00	2619.05
16 Eastern Germany SCG/BAC	26.32	38	18.42	34.21	7.89	7.89	2593.66
17 Eastern Germany SMC	10.00	30	3.33	16.67	6.67	6.67	2594.38
18 Western Germany BAC	14.29	14	0.00	28.57	14.29	14.29	2651.59
19 Western Germany BEC	17.65	17	5.88	29.41	11.76	11.76	2670.34
20 Western France Prissé	0.00	3					3230.44
21 South-Eastern France Treilles	6.90	29	0.00	13.79	6.90	6.90	2923.48
22 Catalonia Epicardial	42.86	7					2944.35

23 Catalonia Late Epicardial	33.33	3	2972.00
24 Spain Basque country	0.00	7	3382.41
25 Portugal coastal Late Neolithic	0.00	3	3934.71
26 Portugal inland Late Neolithic	0.00	6	3819.67

Data S7

Early Neolithic K haplotypes relative to rCRS. Data obtained from the sources in S1 Data and used in Sec. 5.8.1.

Sample ID	Region	mtDNA haplogroup	cal BCE	Haplotype	Polymorphisms relative to rCRS (nucleotide positions 16106-16390)
H25	1 Syria PPNB	K	7400	H01	T16224C T16311C
H4	1 Syria PPNB	K	7400	H02	T16311C
H7	1 Syria PPNB	K	7400	H02	T16311C
R65-14	1 Syria PPNB	K	6975	H03	T16224C T16311C C16366T
R65-1S	1 Syria PPNB	K	6975	H03	T16224C T16311C C16366T
R65-C8-SEB	1 Syria PPNB	K	6975	H03	T16224C T16311C C16366T
Bar8	2 Anatolia	K1a2	6121	H01	T16224C T16311C
I0746	2 Anatolia	K1a or K1a1	6300	H01	T16224C T16311C
I1100	2 Anatolia	K1a or K1a6	6300	H04	T16224C T16311C G16290A
I0727	2 Anatolia	K1a2	6000	H01	T16224C T16311C
I1583	2 Anatolia	K1a2	6300	H01	T16224C T16311C
I1102	2 Anatolia	K1a3a	6300	H01	T16224C T16311C
I0707	2 Anatolia	K1a4	6300	H01	T16224C T16311C
I0724	2 Anatolia	K1a4	6000	H01	T16224C T16311C
I1579	2 Anatolia	K1a-C150T	6300	H05	T16189C T16224C T16311C
I1103	2 Anatolia	K1b1b1	6300	H01	T16224C T16311C
VINJ2	3 Hungary-Croatia Starcevo	K	5700	H08	T16224C C16261T T16311C
VINK1	3 Hungary-Croatia Starcevo	K1a	5700	H01	T16224C T16311C
VINK5	3 Hungary-Croatia Starcevo	K1a	5700	H05	T16189C T16224C T16311C
BAM02	3 Hungary-Croatia Starcevo	K	5735	H07	T16172C T16224C T16311C
BAM04	3 Hungary-Croatia Starcevo	K	5595	H01	T16224C T16311C
BAM16	3 Hungary-Croatia Starcevo	K	5700	H01	T16224C T16311C
M6-116.4	3 Hungary-Croatia Starcevo	K	5700	H06	A16166G T16224C T16311C
BAM07	3 Hungary-Croatia Starcevo	K1	5700	H01	T16224C T16311C
BAM24	3 Hungary-Croatia Starcevo	K1	5700	H01	T16224C T16311C
LGCS4	3 Hungary-Croatia Starcevo	K1a	5700	H05	T16189C T16224C T16311C
BAM09	3 Hungary-Croatia Starcevo	K1a	5700	H05	T16189C T16224C T16311C
BAM19	3 Hungary-Croatia Starcevo	K1a	5700	H01	T16224C T16311C
KAR 10	4 Eastern Germany LBK	K	5030	H01	T16224C T16311C
KAR 54	4 Eastern Germany LBK	K1a	5138	H01	T16224C T16311C
KAR 7	4 Eastern Germany LBK	K1a	5138	H10	T16209C T16224C T16311C
NAU 3	4 Eastern Germany LBK	K1a	5138	H01	T16224C T16311C
KAR 8	4 Eastern Germany LBK	K1b1a	5138	H11	T16224C T16311C G16319A
KAR 55	4 Eastern Germany LBK	K2a5	5138	H01	T16224C T16311C
UWS 3	4 Eastern Germany LBK	K	5138	H01	T16224C T16311C
UWS 2	4 Eastern Germany LBK	K	5138	H09	T16224C T16249C T16311C
deb38	5 Western Germany LBK	K	5138	H01	T16224C T16311C
deb02	5 Western Germany LBK	K	5075	H01	T16224C T16311C
deb29II	5 Western Germany LBK	K	4982	--	Not Determined
deb10	5 Western Germany LBK	K1a	5138	H01	T16224C T16311C

HAL 12	5 Western Germany LBK	K	5138	H01	T16224C T16311C
HAL 18	5 Western Germany LBK	K	5138	H01	T16224C T16311C
HAL 31	5 Western Germany LBK	K	5122	H09	T16224C T16249C T16311C
HAL 20	5 Western Germany LBK	K1a	5138	H01	T16224C T16311C
HAL 9	5 Western Germany LBK	K1a	5138	H01	T16224C T16311C
I0048	5 Western Germany LBK	K1a	5129	H01	T16224C T16311C
I1550	5 Western Germany LBK	K1a2	5138	H01	T16224C T16311C
CSA152223	6 North-Eastern Spain Cardial	K	5390	H01	T16224C T16311C
CB13	6 North-Eastern Spain Cardial	K1a2a	5415	H01	T16224C T16311C
1CH0102	6 North-Eastern Spain Cardial	K	5164	H01	T16224C T16311C
Troc4	6 North-Eastern Spain Cardial	K1a2a	5123	H01	T16224C T16311C
PAT-2E2	7 Spain Navarre	K	4968	H01	T16224C T16311C
CAS-181	7 Spain Navarre	K1a	4933	H01	T16224C T16311C
CAS-202	7 Spain Navarre	K1a	4933	H01	T16224C T16311C
CAS-191	7 Spain Navarre	K1a	4933	H01	T16224C T16311C
Gök5	11 Sweden	K1a5	2951	H12	T16224C T16311C T16362C

Appendix B. Copy of original publications derived from this thesis

Journal of Theoretical Biology 385 (2015) 112–118

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Front propagation speeds of T7 virus mutants

V.L. de Rioja*, J. Fort, N. Isern

Complex Systems Laboratory, Departament de Física, Universitat de Girona, 17071 Girona, Catalonia, Spain

HIGHLIGHTS

- We present a spatial spread model of virus infections.
- The virus–host bacteria interaction has an improvement linked to the delay time.
- Our model is biologically sound and satisfactorily explains the experimental results.
- Some previous models did not consider the diffusive delay yielding too fast speeds.

ARTICLE INFO

Article history:
Received 18 December 2014
Received in revised form
22 July 2015
Accepted 1 August 2015
Available online 21 August 2015

Keywords:
Biophysics
Front propagation
Reaction–diffusion equations
Mathematical model
Virus dynamics

ABSTRACT

We propose a new reaction–diffusion model with an eclipse time to study the spread of viruses on bacterial populations. This new model is both biologically and physically sound, unlike previous ones. We determine important parameter values from experimental data, such as the one-step growth. We verify the proposed model by comparing theoretical and experimental data of the front propagation speed for several T7 virus strains.

1. Introduction

Bacterial viruses or bacteriophages (literally 'eaters of bacteria') infect and replicate within bacteria. Right after their discovery, phages were used as an early form of biotechnology to fight bacterial pathogens. Nowadays, drug-resistant strains for many bacteria have appeared and this has lead to a revived interest in this kind of therapy (Weinbauer, 2004). Moreover, these viruses are among the most common and diverse entities in the biosphere, so it is important to attain a better and more accurate knowledge of their dynamics. Understanding the speed of virus infection fronts is also important in the context of cancer treatment (Wodarz et al., 2012).

It is possible to see with the naked eye how the spreading dynamics of viruses works in a medium of susceptible host bacteria. When a small quantity of phages is inoculated into a tiny, central region of liquid agar with host cells (bacteria in our case), the continuous replication and diffusion of viruses lead to an enlarging dark region, composed of dead cells. Such a region of lysed (i.e. dead) cells, surrounded by unlysed cells, is called a plaque. The growth process starts when a free virus diffuses into a host bacterium, adsorbs on its surface, injects its DNA into it, replicates within and finally (after a certain time) the bacterium dies and expels a new generation of viruses. The progeny viruses diffuse to surrounding host cells, and the cycle repeats again. The propagating front has a well-characterized speed, typically less than a millimeter per hour, which has been measured experimentally, and for which we try below to get a realistic and accurate reaction–diffusion model.

Numerous models of phage plaque enlargement have been proposed. The oldest and simplest one is due to Koch (1964), who suggested that the diffusion speed was proportional to $\sqrt{D/\tau}$, where D is the diffusion coefficient and τ is the phage latent period (i.e., the time during which bacteriophages are inside cells and thus not moving). By incorporating additional kinetic parameters, Yin and McCaskill (1992) constructed a reaction–diffusion system and obtained the speed of traveling-wave solutions. Later You and Yin (1999) supported the previous idea of an existing traveling-wave solution through numerical simulations of the same problem. However the models due to Yin and co-workers (Yin and McCaskill, 1992; You and Yin, 1999) lead, *for parameter values derived from independent experiments*, to speeds much faster than the experimental ones (Yin

* Corresponding author.
E-mail address: victor.lopezd@udg.edu (V.L. de Rioja).

http://dx.doi.org/10.1016/j.jtbi.2015.08.005
0022-5193/

and McCaskill, 1992; Fort, 2002). It was then realized that the delay time or latent period (i.e., the time interval during which a virus is inside a cell and thus does not move) delays virus diffusion, and that this important effect could explain the slowness of the experimental speeds (Fort and Méndez, 2002). By solving the problem numerically, good agreement with experiment was attained (*without fitting any parameter values*) (Fort and Méndez, 2002). However the equations were not fully understood from a biological viewpoint, as we shall explain below. Later Ortega-Cejas et al. (2004) obtained some approximate but explicit formulas for the front speed based on the model in Fort and Méndez (2002). Among more recent models, Amor and Fort (2010) proposed a new improved set of equations which satisfactorily explained the observed speeds of VSV (vesicular stomatitis virus) infections, but still with some terms lacking a clear biological interpretation.

Using various bacteriophage T7 mutants in a growing plaque on *E. coli* host bacteria, Yin (1993) measured experimentally the radial propagation speed for plaques of three mutant T7 virus strains (namely, p001, p005 and the wild type), finding different speeds depending on the type of mutant. These are the experiments that we want to explain.

On the basis of the model for VSV infections (Amor and Fort, 2010), we rewrite the equations carefully, so that they acquire full biological and mathematical meaning, and we apply them to T7 strains. With this new model we obtain a good agreement with the experimental results in Yin (1993), without requiring the use of any free or adjustable parameters.

In this paper, we introduce a new reaction–diffusion set of equations to explain the existing experimental data on the growth of T7 plaques on bacteria. In Section 2 we present the new time-delayed model and we discuss why our modifications are reasonable. Section 3 is devoted to estimations of the necessary parameter values from independent experiments. In Section 4, the results are compared with experimental data for the propagation speed of three strains of the T7 virus, and Section 5 presents a simplification of the model yielding similar results. In Section 6 we compare to other time-delayed models. Finally, Section 7 is devoted to final conclusions, with particular attention to the model features and how the results are improved over previous models.

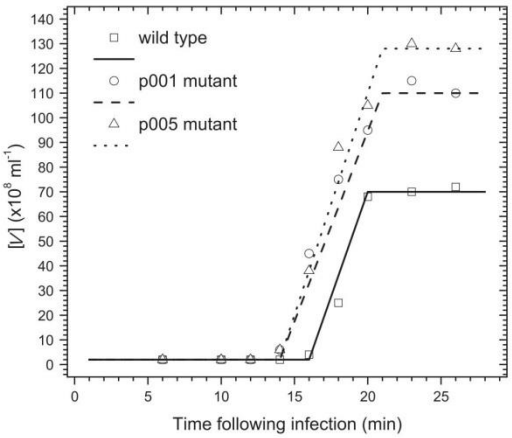


Fig. 1. One-step growth curves of T7 mutants adapted to the model in this paper. Experimental data (□ for the wild T7, ○ for the p001 mutant and △ for the p005 mutant) have been obtained from Fig. 3 in Yin (1993). Full, dashed and dotted lines correspond to the fits for the wild type and p001 and p005 mutants, respectively.

2. Reaction–diffusion model

We model the spatial dynamics of T7 mutants infecting host cells by considering interactions among three species: viruses (V), uninfected bacteria (B) and infected bacteria (I). Those processes can be described schematically as

$$V+B\xrightarrow{k_1}I\xrightarrow{k_2}Y\cdot V, \tag{1}$$

where k_1 is the adsorption rate, k_2 the death rate of infected bacteria, and Y (yield or burst size) is the number of new viruses released per lysed host bacteria. These three parameters (k_1, k_2 and Y) depend on the mutant strain considered.

The experiment on which this theoretical work is focused (Yin, 1993) was conducted in agar (so that host bacteria are immobilized) and cells were initially in the stationary phase, i.e. with bacterial growth and death in balance (so that the number density of live bacteria does not change appreciably before viruses arrive). Viruses can move and adsorb on host bacteria, infecting cells and producing new viruses.

Some previous models have the drawback of assuming logistic dynamics, namely (Fort and Méndez, 2002; Ortega-Cejas et al., 2004; Amor and Fort, 2010)

$$\frac{\partial[I](r,t)}{\partial t}=-k_2[I](r,t)\left\{1-\frac{[I](r,t)}{[I]_{\max}}\right\}, \tag{2}$$

in the absence of uninfected cells (i.e., if all cells are initially infected). $[I]$ in Eq. (2) is the concentration of infected hosts, $[I]_{\max}$ is their maximum concentration, r is the distance from the inoculation point, and t is the time.

Let us define 'free space' as the fraction of space not occupied by infected cells, relative to the maximum possible value that can be occupied by them, i.e. $1-[I](r,t)/[I]_{\max}$. Eq. (2) describes well the one-step growth experiment (see Fig. 1 in Fort and Méndez, 2002) but has no biological meaning. Indeed, it assumes that the death rate of infected cell is proportional not only to the concentration of infected cells $[I]$ (which is reasonable), but also to the free space (term within brackets). Thus, we propose to replace this equation by taking into account the eclipse time τ between adsorption and the onset of the release of the virus progeny. Therefore, in the absence of adsorption we propose to replace Eq. (2) by

$$\frac{\partial[I](r,t)}{\partial t}=-k_2[I](r,t-\tau). \tag{3}$$

Note that we do not assume that all cells die at the same time after infection. That assumption is made in the perfect delay model (Jones et al., 2012) , which makes use of $[V](r,t-\tau)$ $[B](r,t-\tau)$ instead of $k_2[I](r,t-\tau)$ (as we will see in Section 6 in detail). But the perfect delay model disagrees with biological experiments (because one-step experiments do not display a vertical step, see Fig. 1). In contrast, Eq. (3) does not represent a perfect vertical step, but a gradual increase after an eclipse time τ. The model we present below is an alternative to the exponential, non-delayed model [i.e., a term $k_2[I](r,t)$], and the perfect delay model [i.e., a term $[V](r,t-\tau)[B](r,t-\tau)$], and is more realistic than both extreme models.

In the presence of adsorption, the model we propose is thus described by (see Appendix A):

$$\frac{\partial[I](r,t)}{\partial t}=k_1[V](r,t)[B](r,t)-k_2[I](r,t-\tau), \tag{4}$$

$$\frac{\partial[V](r,t)}{\partial t}+\frac{\tau}{2}\frac{\partial^2[V](r,t)}{\partial t^2}=D_{eff}\frac{\partial^2[V](r,t)}{\partial r^2}+F(r,t)-\frac{\tau}{2}k_1[V](r,t)\frac{\partial[B](r,t)}{\partial t}$$

$$-\frac{\tau}{2}k_1[B](r,t)F(r,t)+\frac{\tau}{2}k_2Y\frac{\partial}{\partial t}[I](r,t-\tau)], \tag{5}$$

$$\frac{\partial[B](r,t)}{\partial t}=-k_1[V](r,t)[B](r,t), \tag{6}$$

where $[V]$ and $[B]$ are the concentration of viruses and uninfected bacteria respectively, and D_{eff} is the effective diffusion coefficient of viruses (see next section). Bacteria do not diffuse because they are immobilized by the agar in this experiment. The virus growth function, $F(r,t)$, in Eq. (5) is

$$F(r,t)=-k_1[V](r,t)[B](r,t)+k_2Y[I](r,t-\tau). \tag{7}$$

In this model [Eqs. (4)–(7)], the time derivative $\partial/\partial t$ represents the change of the population number over time and the second space derivative $\partial^2/\partial r^2$ is related to the diffusion through space. Terms proportional to k_1 account for the decay of viruses [Eqs. (5) and (7)] and host bacteria [Eq. (6)] and the creation of infected cells [Eq. (4)], as a result of the infection process (note that these terms are the same as Eqs. (9) and (11)–(13) in Amor and Fort, 2010). Infected cells also decay following their own rate of death k_2 [Eq. (4)], and as shown in Eq. (1), for each dead cell the viruses increase their number Y times [Eqs. (5) and (7)]. The terms proportional to τ in Eq. (5) are second-order corrections (see Appendix A), they were applied already in Amor and Fort (2010), and they take care of the time delay due to the fact that viruses spend a time τ inside cells before the new generation disperses away (Fort and Méndez, 2002). As mentioned above, the main drawback of Amor and Fort (2010) is studying the death of the infected cells from a logistic equation, which has no biological sense.

Therefore, here we present a new model with two main effects: (i) the second-order correction that has been shown to be fundamental to describe time-delayed biological fronts (Fort and Méndez, 1999a, 2002; Amor and Fort, 2010) (i.e., the terms proportional to τ in Eq. (5)), and also (ii) a biologically meaningful description of the death process, Eq. (3) (instead of logistic growth dynamics, Eq. (2)).

Other authors have also described the death process through including an eclipse time with terms proportional to concentrations at $t-\tau$, rather than a logistic function (Jones et al., 2012; Gourley and Kuang, 2005). However, those models do not include any second-order terms, i.e. any diffusive delay (effect (i) in the previous paragraph), which is necessary to take proper account of the fact that viruses do not move during a time interval τ, because they are inside the infected cells. The death of infected cells is also described in Jones et al. (2012) and Gourley and Kuang (2005) differently than in our model (in Section 6 we discuss this in more detail and compare the models and experiments).

We introduce dimensionless variables to simplify the analysis. Let B_0 be the initial concentration of bacteria, then $\overline{B}\equiv[B]/B_0$, $\overline{V}\equiv[V]/B_0$, $\overline{I}\equiv[I]/B_0$, $\overline{t}\equiv k_2t$ and $\overline{r}\equiv r\sqrt{k_2/D_{eff}}$ are the new dimensionless variables, and $\overline{\tau}\equiv k_2\tau$ and $\kappa\equiv k_1B_0/k_2$ the new dimensionless parameters. The aim is to find the speed of traveling-wave solutions which satisfies the set of differential equations (4)–(6). These become single-variable differential equations by using the co-moving coordinate $\overline{z}\equiv\overline{r}-\overline{c}\overline{t}$. $\overline{c}$ (positive) is the dimensionless wave front speed and is related to the dimensional speed c by $\overline{c}=c/\sqrt{k_2D_{eff}}$. Following previous work (Yin and McCaskill, 1992; Fort and Méndez, 2002; Amor and Fort, 2010), we assume that the concentrations at the leading edge of the propagation front ($\overline{z}\to\infty$) are $(\overline{V},\overline{B},\overline{I})=(\epsilon_V,1-\epsilon_B,\epsilon_I)\simeq(0,1,0)$, where $\vec{\epsilon}=(\epsilon_V,\epsilon_B,\epsilon_I)=\vec{\epsilon_0}\cdot\exp(-\lambda\overline{z})$. For non-trivial solutions to exist, the determinant of the matrix corresponding to the linearized model must be zero. This leads us to the following characteristic equation:

$$\left(1-\frac{\overline{\tau}}{2}\overline{c}^2\right)\overline{c}\lambda^3+\left[e^{-\lambda\overline{c}\overline{\tau}}-\overline{c}^2\left(1+\frac{\overline{\tau}}{2}e^{-\lambda\overline{c}\overline{\tau}}\right)\right]\lambda^2+\left[\kappa\left(\frac{\overline{\tau}}{2}\kappa-1+\frac{\overline{\tau}}{2}Ye^{-\lambda\overline{c}\overline{\tau}}\right)-e^{-\lambda\overline{c}\overline{\tau}}\right]\overline{c}\lambda+\kappa e^{-\lambda\overline{c}\overline{\tau}}\left[\frac{\overline{\tau}}{2}\kappa-1-Y\left(\frac{\overline{\tau}}{2}\kappa-1\right)\right]=0. \tag{8}$$

It is known that, according to marginal stability analysis (Ebert and van Saarloos, 2000), the propagation front moves with the minimum possible speed. Therefore,

$$\overline{c}=\min_{\lambda>0}[\overline{c}(\lambda)], \tag{9}$$

where $\overline{c}(\lambda)$ is given implicitly by Eq. (8).

3. Parameter values

We have a new time-delayed model which depends on various parameters. It is necessary to estimate their values from experiments different from the front-speed experiments that we want to explain. The front propagation speed depends on the viral diffusivity D_{eff}, the average yield Y, the kinetic parameters k_1 and k_2, the host concentration B_0 and the eclipse time τ. Since we aim to explain the experimental data in Yin (1993), these parameters must be determined for strains of the T7 virus and *E. coli* bacteria.

Yin and co-workers noted that the diffusion coefficient D of viruses in agar must be corrected by the fact that host bacteria adsorb the viruses, and this leads to more tortuous paths for the viruses at high bacterial concentrations. As noted in previous work, the effective coefficient D_{eff} is therefore given by Fricke's law (Fort and Méndez, 2002),

$$D_{eff}=\frac{1-f}{1+\frac{f}{x}}D, \tag{10}$$

where f is the initial concentration of bacteria relative to its maximum, i.e. $f=B_0/B_{\max}$, and x stands as an approximation of the cells' shape. For spherical particles $x=2$, while for *E. coli* it is more accurate to use $x=1.67$ (Fort and Méndez, 2002). Note that the diffusivity coefficient D corresponds to viruses moving through agar in the absence of bacteria ($f=0$). T7 viruses are very similar to phage P22 in shape and size, thus we use the corresponding value $D=4\times10^{-8}$ cm^2/s (Fort and Méndez, 2002).

The rate of adsorption of viruses, k_1, was estimated from a separate experiment conducted in KCN, a substance that prevents viruses from reproducing. We have only one experimental value for the T7 virus, $k_1=(1.29\pm0.59)\times10^{-9}$ ml/min (Fort and Méndez, 2002), corresponding to the wild strains. For the other mutants, we have not been able to find any reliable experimental value, thus we will use the same value of k_1 for all three strains. We will return to discuss this parameter in the next section.

Finally, the parameters τ, Y and k_2 are obtained from the so-called one-step growth experiments. They consist in measuring the concentration of viruses as a function of time for a given initial, homogeneous population of infected bacteria. Depending on the T7 mutant, the curves are different (see Fig. 1) and so will be the parameter values. Fig. 1 allows us to obtain the necessary information from each mutant to estimate its value of τ, Y and k_2, as we next explain.

The eclipse phase of the one-step growth (Fig. 1) corresponds to the stage between adsorption ($t=0$) and the first release of viruses (i.e., the beginning of the rise in virus density). This interval of time (the eclipse time) is 16 min for the wild type and 14 min for the p001 and p005 mutants. Note that if we used higher values for τ, e.g. $\tau=18$ min for the wild strain, for $t=17$ min

we would have $\partial[I](r,t)/\partial t = 0$ according to Eq. (3), whereas we must have $\partial[I](r,t)/\partial t \neq 0$ according to Fig. 1.

The average burst sizes Y differ significantly for the three mutants. They can be calculated as the quotient between the maximum and the initial concentration of viruses, i.e. $Y = (V_{max})/V_0$, where according to Fig. 1 $V_0 = 2 \times 10^8$ ml^{-1} for the three kinds of mutants. Inserting the data in Fig. 1, we obtain the yields Y=34.5 for the wild type, Y=56.5 for the p001mutant, and Y=65 for the p005 mutant. As we shall see, these higher productivities of new generations for the two mutants result in faster infections (relative to the wild type). The three yields above have been obtained for cells in agar-immobilized microcolonies containing many cells. As noted by Yin and McCaskill (1992), such yields are substantially lower than the typical yield for an isolated cell under optimal conditions ($Y \approx 200$). Yin and McCaskill (1992) suggested that this difference may be due to a number of factors, such as inherently lower yields per cell when immobilized in agar, premature lysis or inhibition due to the death of adjacent cells, high multiplicities of adsorption required for host infection, readsorption of newly released viruses on cell fragments, etc. If we used the yield for an isolated cell, we would have to incorporate additional terms to include other possible kinds of death and interactions in our mathematical model. However, the measured experimental values of the burst size quoted above (for cells in agar-immobilized microcolonies containing many cells) implicitly include these possible interactions.

The rate of death of infected bacteria k_2 may be understood as the reproduction of viruses, because viruses replicate as bacteria die. For $t < \tau$ there are no new viruses (see Fig. 1), so no infected cells have died yet and thus $[I](t) = I_0$. For $t \geq \tau$ Eq. (3) yields $d[I] = -k_2 I_0 dt$. Because each infected cell produces Y viruses, $d[V] = -Yd[I] = k_2 Y I_0 dt = k_2 V_{max} dt$. Therefore, the slope of each straight line in Fig. 1 is $k_2 V_{max}$, and $k_2 = V_{max} - V_0 / \Delta t \cdot V_{max} \approx 1/\Delta t$, where Δt is the time interval during which $[V]$ increases, also known as the rise period $1/k_2$. It is straightforward to estimate the values of k_2 from the figure, and they turn out to be 1/4 min^{-1} for the wild type and 1/6 min^{-1} for the p001 and p005 mutants.

It is important to remember that all of these parameters are known a priori, thus we do not use any free or adjustable parameters in our predictions.

Accordingly to Fig. 1, the average latent period is $\tau + 1/2k_2$, where τ is the eclipse time (the factor 1/2 is due to the fact that, after a cell is infected, the first viruses leave it after a time interval τ, and the last viruses leave it after a time interval τ+1/k2, see Fig. 1).

For clarity, we mention that when all infected cells have died, no more viruses are produced (Fig. 1, right side) and Eq. (3) obviously breaks down. Thus the general evolution equation we propose is

$$\frac{\partial[I](r,t)}{\partial t} = \begin{cases} -k_2[I](r,t-\tau) & \text{if} \quad [V] < V_{max} \\ 0 & \text{if} \quad [V] = V_{max} \end{cases} \tag{11}$$

where the second line is analogous to some approaches to single-species systems (see Eq. (9) in Fort et al., 2007). However this point is, in fact, unnecessary for the purposes of the present paper because the front speed is computed at the leading edge of the infection front, where $[V] \simeq 0$ (Section II). Obviously, in Eq. (11) the condition $[V] < V_{max}$ is equivalent to $[I] \neq 0$, and the condition $[V] = V_{max}$ is equivalent to $[I] = 0$.

4. Theory versus experiment

In this section we study the spatial dynamics of different T7 virus strains. The experimental data (black squares in Fig. 2) and their error bars were obtained in Yin (1993) for plaques where the concentration of nutrient was 10 g/l, which corresponds to f=0.2 (see Yin and McCaskill, 1992, pp. 1543–1544) and $B_{max} = 10^7$ ml^{-1} (see Yin and McCaskill, 1992, Fig. 3a) thus $B_0 = 2 \times 10^6$ ml^{-1}.

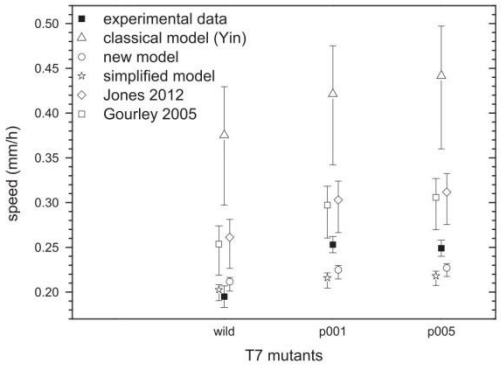


Fig. 2. Front propagation speeds for T7 mutants (wild, p001 and p005). Black squares refer to experimental data and white symbols to the theoretical models: triangles for the classical Yin et al. model, circles for the new model, stars for the simplified model explained in Section 5, rhombuses for the model by Jones et al. (2012), and white squares for the model by Gourley and Kuang (2005) both from Section 4.

The theoretical results will be calculated below with the parameters Y, k_2 and τ for each strain extracted from Fig. 1 (as detailed in Section 3), and the mean values of k_1 and D_{eff}. Because the value of k_1 is substantially more uncertain than those of other parameters, the corresponding error bars are obtained from the experimental range of k_1, namely $k_1 = (1.29 \pm 0.59) \times 10^{-9}$ ml/min (Fort and Méndez, 2002).

The classical approach with no delay or eclipse time, due to Yin and McCaskill (triangles in Fig. 2), predicts speeds much faster than the experimental ones (black squares). This model by Yin and McCaskill (1992) (with $k_{-1} = 0$, as noted in Yin and Amn, 1994) is the same as our model [Eqs. (4)–(7)] with $\tau = 0$, i.e.

$$\frac{\partial[I](r,t)}{\partial t} = k_1[V](r,t)[B](r,t) - k_2[I](r,t), \tag{12}$$

$$\frac{\partial[V](r,t)}{\partial t} = D_{eff}\frac{\partial^2[V](r,t)}{\partial r^2} - k_1[V](r,t)[B](r,t)k_2[I](r,t) \tag{13}$$

$$\frac{\partial[B](r,t)}{\partial t} = -k_1[V](r,t)[B](r,t), \tag{14}$$

The new model introduced in this paper (circles in Fig. 2) agrees better to the experimental data than the classical model Yin and McCaskill, for all three mutants. This improvement is clearly visible in Fig. 2, where we see that the results from the new model lie much closer to the experimental data than the classical model (Yin and McCaskill, 1992). If we calculate the errors of the models versus the experimental data, the classical model by Yin and McCaskill has an average error of 75%, compared to only 10% for the new model presented here.

5. Simplified mathematical model

Our new model, Eqs. (4)–(7), yields a rather complex characteristic equation, Eq. (8), from which we compute the front speeds. In this section we derive a simplified expression leading to similar results. We proceed by removing each term and evaluating its contribution to the front speed, in order to ultimately keep only those terms that have a major contribution on the model results.

In this way, it is easy to see that all of the terms in Eqs. (4) and (6) are important to achieve a good result, but some terms in Eq. (5) are not. Hence, we just modify this equation.

On one hand, the expansion of $F(r,t)$ to second-order (the three last terms in Eq. (5)) introduces a small change on the results. We can neglect all reaction terms proportional to τ in this equation.

On the other hand, if we understand the right side of Eq. (5) as the diffusion term, plus the reaction term (plus second-order approximations), we can also neglect the adsorption of virus into bacteria, i.e. the term with k_1 in Eq. (7). Diffusion and creation of new viruses are thus the terms with major contributions to the front speed.

In this simplified model we can therefore replace Eq. (5) in our set by

$$\frac{\partial[V](r,t)}{\partial t}+\frac{\tau}{2}\frac{\partial^2[V](r,t)}{\partial t^2} = D_{eff}\frac{\partial^2[V](r,t)}{\partial r^2}+k_2 Y[I](r,t-\tau). \quad (15)$$

Considering now the set composed by Eqs. (4), (6) and (15) we obtain a new characteristic equation,

$$\left[\lambda\bar{c}+\lambda^2\left(\frac{\bar{\tau}}{2}\bar{c}^2-1\right)\right]\left(\lambda\bar{c}+\mathrm{e}^{-\lambda\bar{c}\,\bar{\tau}}\right)-\kappa Y\mathrm{e}^{-\lambda\bar{c}\,\bar{\tau}}=0, \quad (16)$$

much simpler than the previous equation (8). The results of this model are shown as stars in Fig. 2. As it can be seen, the front speeds of the simplified model (stars) are always slightly slower than those found by the main model (circles). But the difference between the two models is only about 4% in all three cases. By comparing with the experimental data (black squares in Fig. 2), we see that the simplified model in this section (stars, Eq. (16)) is still much better than the classical one (triangles) in spite of being much simpler than the complete model in Section 2 (circles, Eq. (8)).

6. Comparison to other time-delayed models

Some other authors have also described the death process by considering concentrations at $t-\tau$, rather than a logistic function (Jones et al., 2012; Gourley and Kuang, 2005). However, as mentioned above, those models do not include the diffusive delay (i.e., second-order corrections), which is necessary because viruses do not diffuse when they are inside the infected cells. Another difference between our model and that in Jones et al. (2012) is that the term $k_2 I(r,t-\tau)$ in our model is replaced by $k_1 B(r,t-\tau)V(r,t-\tau)$. From a conceptual point of view, in our model the infected cells present at the system at time $t-\tau$ begin to die at time t, and do so gradually thereafter (with rate k_2). Thus not all cells die exactly at time t in our model, in agreement with the experimental data (Fig. 1). In contrast, according to the model in Jones et al. (2012) all cells infected at time $t-\tau$ die exactly at time t, thus in the one-step experiment their model predicts a perfect step-like result, in disagreement with experimental data (Fig. 1). Thus we expect the model by Jones et al. (2012) to yield faster speeds than our model for two reasons: (i) they neglect the diffusive delay and (ii) they neglect the fact that the death of some cells takes longer than τ after infection. Replacing D by D_{eff} (as explained in Section 3), the model by Jones et al. (2012) is (see Eqs. (2.2))

$$\frac{\partial[I](r,t)}{\partial t} = k_1[V](r,t)[B](r,t) - k_1[V](r,t-\tau)[B](r,t-\tau), \quad (17)$$

$$\frac{\partial[V](r,t)}{\partial t} = D_{eff}\frac{\partial^2[V](r,t)}{\partial r^2}-k_1[V](r,t)[B](r,t) + Yk_1[V](r,t-\tau)[B](r,t-\tau), \quad (18)$$

$$\frac{\partial[B](r,t)}{\partial t} = -k_1[V](r,t)[B](r,t). \quad (19)$$

Note that this is the same as the model by Yin et al. [Eqs. (12)–(14)], with $k_2 I(r,t)$ replaced by $k_1 B(r,t-\tau)V(r,t-\tau)$. By following again the same method as in Section 2, we find that the characteristic equation for the model due to Jones et al. (2012) is

$$\lambda^2-\lambda\bar{c}+\kappa\left(Y\mathrm{e}^{-\lambda\bar{c}\,\bar{\tau}}-1\right)=0. \quad (20)$$

Note that, in fact, the equation for $\partial[I](r,t)/\partial t$ above is not necessary to compute this speed, since $[I]$ does not appear in the other two equations of the model by Jones et al. (2012).

In Fig. 2 (rhombuses) we have also included the predictions of the model by Jones et al., for the same parameter values used in our model. We see in Fig. 2 that their model (Jones et al., 2012) predicts faster speeds than our model, as expected. Moreover, they are faster than the experimental speeds. For the wild strain, our model is consistent with the experimental range. For the mutants p001 and p005, the mean speeds predicted by our model are also closer to the experimental means (although the error bars are larger for the model by Jones et al. (2012), because the speed depends strongly on k_1).

There is one more time-delayed model of virus front spread, due to Gourley and Kuang (2005). It is very similar to that by Jones et al. (2012), discussed above, but it assumes an additional, natural death process only for *infected* cells (with rate μ_I and unrelated to virus infection) that decreases the number density of *infected* cells after time τ by a factor $e^{-\mu_I\tau}$ (Gourley and Kuang, 2005). Although no biological reason was given in Gourley and Kuang (2005) why an additional death process might affect only the infected cells (and not the uninfected ones), for completeness we next explore whether this model by Gourley and Kuang (2005) changes the results of the model by Jones et al. (2012) or not. Since this model by Gourley and Kuang (2005) includes an additional death process for the infected cells, intuitively we expect that it could yield slower speeds than the model due to Jones et al. (2012). For the experimental conditions corresponding to the speeds that we analyze in the present paper (Fig. 2), the model proposed by Gourley and Kuang (2005) is (see Eqs. (1.1), (2.1) and (4.1))

$$\frac{\partial[I](r,t)}{\partial t} = k_1[V](r,t)[B](r,t) - \mathrm{e}^{-\mu_I\tau}k_1[V](r,t-\tau)[B](r,t-\tau), \quad (21)$$

$$\frac{\partial[V](r,t)}{\partial t} = D_{eff}\frac{\partial^2[V](r,t)}{\partial r^2}-k_1[V](r,t)[B](r,t) + Y\mathrm{e}^{-\mu_I\tau}k_1[V](r,t-\tau)[B](r,t-\tau), \quad (22)$$

$$\frac{\partial[B](r,t)}{\partial t} = -k_1[V](r,t)[B](r,t), \quad (23)$$

where we have neglected cell reproduction because in the experiments we want to explain, the cells were in the stationary growth phase before the arrival of viruses (as explained in 2). We have also neglected virus death because it is negligible (de Paepe and Taddei, 2006). We do not include diffusion of uninfected or infected cells because bacteria are immobilized in agar in these experiments (as mentioned in Section 2). By following again the same method, the characteristic equation in the model due to Gourley et al. is

$$\lambda^2-\lambda\bar{c}+\kappa\left(Y\mathrm{e}^{-\lambda\bar{c}\,\bar{\tau}}\mathrm{e}^{-\mu_I\tau}-1\right)=0. \quad (24)$$

Again, in fact the equation for $\partial[I](r,t)/\partial t$ above is not necessary to compute this speed, since $[I]$ does not appear in the other two equations of the set. In Fig. 2 (plotted as white squares) we have also included the predictions of this model by Gourley and Kuang

(2005) using the experimental value $\mu_I = 0.4\ \mathrm{h}^{-1}$ (from Fig. 7 in Zobell and Cobet, 1962). It is seen that its predictions are slower (as expected) but almost the same as those of the model by Jones et al. (2012). The speeds from both models are faster than the experimental ones.

Finally, it is worth to note that, in situations where infected cells exit that class due to some other form of interaction, it would be necessary to modify our model. For example, for an additional, natural death process with exponential dynamics for infected cells, the right-hand side in Eq. (4) would include an additional term $-\mu_I[I](r,t)$ and Eq. (5) should be modified accordingly.

7. Conclusions

We have proposed a new reaction–diffusion model with an eclipse time, that satisfactorily explains the experimental results of T7 virus plaques on *E. coli*. This improvement over previous models has been attained by means of the careful modification of one of the evolution equations, which lacked biological significance.

Indeed, some previous models (Fort and Méndez, 2002; Ortega-Cejas et al., 2004; Amor and Fort, 2010) assumed that the death rate of infected cells is proportional not only to their density, but also to the free space [Eq. (2)], which is not biologically reasonable. In contrast, the new model assumes that the death rate is proportional only to the density of infected cells (which begin to die after a time lag τ, corresponding to the eclipse phase of Fig. 1) [Eq. (3)]. Thus our new model is more reasonable biologically. Moreover, our new model agrees reasonably well with experimental data, in contrast to the classical model without delay or eclipse time due to Yin and McCaskill (1992) and You and Yin (1999). It is important to stress that Yin and co-workers already noted that their model was too fast for realistic parameter values, and only by fitting three parameters could it yield sufficiently slow speeds to agree with the experimental ones (Fig. 3 in Yin and McCaskill, 1992). In contrast, here we have not fitted any parameter but used realistic values, i.e. all parameter values we have applied have been obtained from independent experiments.

Other authors took into account the role of the eclipse or delay time τ, but only in the reactive and not in the diffusive process (Jones et al., 2012; Gourley and Kuang, 2005), and they assumed the same eclipse time for all viruses. Those models yield faster speeds than the experimental ones. Also, we stress the importance of using realistic terms to modelize the interactions, e.g. the death process of infected cells (i.e. the release of viral progeny).

Since the propagation of viruses is an active field of study in biophysics and medicine, having an underlying theory that is both mathematically and biologically sound is of special relevance. Furthermore, we have found that the results agree with experiments.

By means of the detailed analysis of a simple mathematical model, we have aimed to demonstrate that such physical models are able to explain the spatial dynamics of virus infections. Certainly, in order to have a more comprehensive understanding of the problem, extensive data gathering for several viruses and environments should be undertaken.

Acknowledgments

This work was partially funded by ICREA (Academia award) and the MINECO (projects SimulPast-CSD2010-00034, FIS-2009-13050 and FIS-2012-31307).

Appendix A. Time-delayed diffusion

In order to make this paper as self-contained as possible, here we include a brief derivation of Eq. (5). The derivation below (see Fort and Méndez, 1999a; Isern and Fort, 2009 for details) was originally proposed for human populations (Fort and Méndez, 1999a) and later applied to viruses (Fort and Méndez, 2002; Ortega-Cejas et al., 2004; Amor and Fort, 2010).

During a time interval equal to the eclipse time τ (estimated from Fig. 1 in our case), the virus concentration changes both due to the reactive processes (1) and to dispersal. We first calculate the former change by using a Taylor series,

$$[[V](x,y,t+\tau)-[V](x,y,t)]_r = \tau\left.\frac{\partial[V]}{\partial t}\right|_r + \frac{\tau^2}{2}\left.\frac{\partial^2[V]}{\partial t^2}\right|_r + \cdots = \tau F + \frac{\tau^2}{2}\left.\frac{\partial[V]}{\partial t}\right|_r + \cdots \tag{25}$$

where the subindex r denotes reactive processes, and $F([V]) \equiv \partial[V]/\partial t|_r$ is given by Eq. (7) according to the corresponding experiments (see Section 2 and Fort and Méndez, 2002).

Secondly, the change due to dispersal can be calculated by defining the dispersal kernel $\phi(\Delta_x,\Delta_y)$ as the probability per unit area that a virus initially placed at $(x+\Delta_x, y+\Delta_y)$ has moved to (x,y) after a time interval τ. Thus,

$$[[V](x,y,t+\tau)-[V](x,y,t)]_d = \int\int [V](x+\Delta_x,y+\Delta_y,t)\phi(\Delta_x,\Delta_y)\mathrm{d}\Delta_x\mathrm{d}\Delta_y - [V](x,y,t). \tag{26}$$

In a system involving both reactive and dispersal processes, we add up their contributions

$$[V](x,y,t+\tau)-[V](x,y,t) = \int\int [V](x+\Delta_x,y+\Delta_y,t)\phi(\Delta_x,\Delta_y)\mathrm{d}\Delta_x\mathrm{d}\Delta_y - [V](x,y,t) + \tau F + \frac{\tau^2}{2}\left.\frac{\partial F}{\partial t}\right|_r + \cdots \tag{27}$$

Assuming that the kernel is isotropic, i.e., $\phi(\Delta_x,\Delta_y) = \phi(\Delta)$, with $\Delta = \sqrt{\Delta_x^2+\Delta_y^2}$, and Taylor-expanding Eq. (27) up to second order in time and space,

$$\frac{\partial[V]}{\partial t} + \frac{\tau}{2}\frac{\partial^2[V]}{\partial t^2} = D\left(\frac{\partial^2[V]}{\partial x^2} + \frac{\partial^2[V]}{\partial y^2}\right) + F + \frac{\tau\partial F}{2\partial t}\bigg|_r, \tag{28}$$

where $D = \langle\Delta^2\rangle/4\tau = \langle\Delta_x^2\rangle/2\tau = \langle\Delta_y^2\rangle/2\tau$ is the diffusion coefficient.

For large distances $r = \sqrt{x^2+y^2}$ from the inoculation point of viruses ($(x,y) = (0,0)$), $\partial^2[V]/\partial x^2 + \partial^2[V]/\partial y^2 \simeq \partial^2[V]/\partial r^2$ and Eq. (28) is the same as Eq. (5), with F given by Eq. (7) and D replaced by D_{eff} (the reason for the latter change is explained in Section 3). Thus the terms proportional to τ in Eq. (5) arise simply from a second-order Taylor expansion. If the role of the eclipse time is neglected ($\tau \simeq 0$), Eq. (28) reduces to the non-delayed or classical model used by Yin and co-workers (Yin and McCaskill, 1992; You and Yin, 1999), namely (see Eq. (13))

$$\frac{\partial[V]}{\partial t} = D\left(\frac{\partial^2[V]}{\partial x^2} + \frac{\partial^2[V]}{\partial y^2}\right) + F. \tag{29}$$

In general, adding up the reactive and diffusive contributions (as done in Eq. (27)) may not be exact (Fort et al., 2007; Isern et al., 2008; Fort and Pujol, 2008; Fort, 2012; Amor and Fort, 2014; Méndez et al., 2014) and this point is taken into account by the so-called sequential or cohabitation models (see especially Fort et al., 2007, Fig. 1 of Isern et al., 2008 and Fig. 17 of Fort and Pujol, 2008). However, for virus infections cohabitation models yield almost the same results as non-cohabitation (or additive) models (Amor and Fort, 2014). Thus in the present paper, we do not take the

cohabitation effect into account for mathematical simplicity (the predicted speeds in Fig. 2 would be the same, so there is no need to use more complicated equations). Let us mention that, in contrast to virus infections, for human waves of advance the cohabitation effect is not negligible (and a more important effect still is due to the shape of dispersal kernels) (Isern et al., 2008; Fort and Soc, 2015). Such more precise models lead to the ballistic speed for fast reproduction (Fort et al., 2010; Fort, 2012), as they should (Fort et al., 2010; Fort, 2012; Méndez et al., 2014). However, for virus infections those corrections are not necessary. In conclusion, the reaction–diffusion equation (5) has the microscopic derivation above and recent criticisms (Méndez et al., 2014) are irrelevant. For $\tau \simeq 0$ and $F = 0$, this also provides a valid derivation of Fickian diffusion (Eq. (29) for $F=0$). It is very important to stress that mathematical arguments (Méndez et al., 2014) are not enough to establish whether a given equation is valid or not, because this depends on the system considered, and must thus be checked by using reactive functions, parameter values and initial conditions appropriate to the experimental setup (for example, to describe the growth of virus plaques, a model with only a pure death process is not realistic, and therefore irrelevant). As another example of this, reaction–diffusion with Fickian diffusion [Eq. (29)] can be applied if the delay time is negligible, which may be justified for some biological species but not for viruses. This is clearly seen in Fig. 2, by comparing our model to the classical or non-delayed one (12)–(14) used by Yin and McCaskill (1992) and You and Yin (1999), which is based on Eq. (29). At the other extreme, the second-order approximation would obviously fail for large τ (Fort and Méndez, 1999b, 2002; Méndez et al., 2014) and, if this happened, additional terms in the Taylor expansions above would be necessary. However, this is not our case (Appendix B). Finally, Fickian diffusion [Eq. (29) with $F=0$] can be applied if the diffusive delay time is sufficiently small, and is useful in many situations (not in our case). Thus parameter values must be examined to choose the appropriate equation for each experiment. Mathematical arguments are not enough, because an equation may be useful to describe some experiments but not others.

For completeness, in Appendix B we extend the derivation above to infinite order and find that the results are similar to those above and in the main paper (second order).

Appendix B. Full time-delayed equation

As shown in Appendix Appendix A, Eq. (5) is in fact an approximation, because it includes only terms up to second order from the Taylor expansions. The virus density $[V]$ rapidly changes on a scale of time smaller than $\tau \approx 15$ min (because the increases in Fig. 1 take 6 min or less). This could therefore lead to errors in the front speeds obtained in Sections 4 and 5. In this appendix we prove that this is not a problem by considering the full time-delayed equation (see Fort and Méndez (1999b), Eqs. (16) and (21)),

$$\sum_{n=1}^{\infty} \frac{\tau^n}{n!} \frac{\partial^n [V](r,t)}{\partial t^n} = \sum_{n=1}^{\infty} \frac{(2D_{eff}\tau)^n}{(2n)!} \frac{\partial^{2n}[V](r,t)}{\partial r^{2n}} + \sum_{n=1}^{\infty} \frac{\tau^n}{n!} \frac{\partial^{n-1} F(r,t)}{\partial t^{n-1}}\Big|_r, \tag{30}$$

instead of its approximation, Eq. (5), together with Eq. (6) and our new Eq. (4). Then, repeating the same steps as in Section 2 we get the following characteristic equation which replaces Eq. (8)),

$$\left(e^{\lambda \bar{c}\bar{\tau}} - \cosh\left(\lambda\sqrt{2\bar{\tau}}\right) - e^{-\kappa\bar{\tau}} + 1\right)\left(\bar{c}\lambda + e^{-\lambda\bar{c}\bar{\tau}}\right) = \frac{\kappa Y}{\lambda\bar{c}+\kappa}\left(1 - e^{-\kappa\bar{\tau}-\lambda\bar{c}\bar{\tau}}\right). \tag{31}$$

Repeating the calculations leading to Fig. 2, but using Eq. (31), we obtain that the differences are very small. Indeed, the error between the second-order approximation and full time-delay equation is lower than 3% for the three strains of the T7 virus. Thus, the use of the second-order approximation in Sections 2–5 is valid.

References

Amor, D.R., Fort, J., 2010. Phys. Rev. E 82, 061905.
Amor, D.R., Fort, J., 2014. Physica A 416, 611.
de Paepe, M., Taddei, F., 2006. PLoS Biol. 4, e193.
Ebert, U., van Saarloos, W., 2000. Physica D 146, 1.
Fort, J., Méndez, V., 1999a. Phys. Rev. Lett. 82, 867.
Fort, J., Méndez, V., 1999b. Phys. Rev. E 60, 5894.
Fort, J., Méndez, V., 2002. Phys. Rev. Lett. 89, 178101.
Fort, J., Pujol, T., 2008. Rep. Progr. Phys. 71, 086001.
Fort, J., Soc., J.R., 2015. Interface 12, 20150166.
Fort, J., Pérez-Losada, J., Isern, N., 2007. Phys. Rev. E 76, 031913.
Fort, J., Pérez-Losada, J., Suñol, J.J., Escoda, L., Massaneda, J., 2010. New J. Phys. 10, 043045.
Fort, J., 2002. J. Theor. Biol. 214, 515.
Fort, J., 2012. Pro. Natl. Acad. Sci. 109, 18669.
Gourley, S.A., Kuang, Y., 2005. SIAM J. Appl. Math. 65, 550.
Isern, N., Fort, J., 2009. Phys. Rev. E 80, 057103.
Isern, N., Fort, J., Pérez-Losada, J., 2008. J. Stat. Mech.: Theor. Exp., P10012.
Jones, D.A., Smith, H.L., Thieme, H.R., Röst, G., 2012. SIAM J. Appl. Math. 72, 670.
Koch, A.L., 1964. J. Theor. Biol. 6, 413.
Méndez, V., Campos, D., Horsthemke, W., 2014. Phys. Rev. E 90, 042114.
Ortega-Cejas, V., Fort, J., Méndez, V., Campos, D., 2004. Phys. Rev. E 69, 031909.
Weinbauer, M.G., 2004. FEMS Microbiol. Rev. 28, 127.
Wodarz, D., Hofacre, A., Lau, J.W., Sun, Z., Fan, H., et al., 2012. PLoS Comput. Biol. 8, 1002547.
Yin, J., 1994. Amn. N. Y. Acad. Sci. 745, 399.
Yin, J., McCaskill, J.S., 1992. Biophys. J. 06, 1540.
Yin, J., 1993. J. Bacteriol. 175, 1272.
You, L., Yin, J., 1999. J. Theor. Biol. 200, 365.
Zobell, C.E., Cobet, A.B., 1962. J. Bacteriol. 84, 1228.

Rioja *et al. Biology Direct* (2016) 11:1
DOI 10.1186/s13062-015-0100-7

Biology Direct

RESEARCH Open Access

A mathematical approach to virus therapy of glioblastomas

Victor Lopez de Rioja*, Neus Isern and Joaquim Fort

Abstract

Background: It is widely believed that the treatment of glioblastomas (GBM) could benefit from oncolytic virus therapy. Clinical research has shown that Vesicular Stomatitis Virus (VSV) has strong oncolytic properties. In addition, mathematical models of virus treatment of tumors have been developed in recent years. Some experiments in vitro and in vivo have been done and shown promising results, but have been never compared quantitatively with mathematical models. We use in vitro data of this virus applied to glioblastoma.

Results: We describe three increasingly realistic mathematical models for the VSV-GBM in vitro experiment with progressive incorporation of time-delay effects. For the virus dynamics, we obtain results consistent with the in vitro experimental speed data only when applying the more complex and comprehensive model, with time-delay effects both in the reactive and diffusive terms. The tumor speed is given by the minimum of a very simple function that nonetheless yields results within the experimental measured range.

Conclusions: We have improved a previous model with new ideas and carefully incorporated concepts from experimental results. We have shown that the delay time τ is the crucial parameter in this kind of models. We have demonstrated that our new model can satisfactorily predict the front speed for the lytic action of oncolytic VSV on glioblastoma observed in vitro. We provide a basis that can be applied in the near future to realistically simulate in vivo virus treatments of several cancers.

Reviewers: This article was reviewed by Yang Kuang and Georg Luebeck. For the full reviews, please go to the Reviewers' comments section.

Keywords: Biophysics, Oncolytic virus, VSV, Glioblastoma, Front propagation speed, Reaction-diffusion equations

Background

Since early last century, viruses have been studied as experimental agents for cancer treatment. The medical interest in the field has fluctuated during this period, reaching a fever pitch in the past two decades. It was in the early 1990s, when recombinant DNA technology became standard, that virus engineering could provide scientific furtherance to virotherapy. Then, oncolytic viruses appeared to be a treatment of tremendous potential and scientists started manipulating them to target cancerous cells more specifically. This culminated in the first marketing approval of an oncolytic virus, granted by the Chinese government in November of 2005 [1]. Very recently, improvements in patient survival have led to endorsements of other oncolytic virus in Europe and the US [2]. In parallel, mathematical models of virus treatment of tumors have been developed [3–5]. However, even with this new ability to engineer viral genomes, a realistic therapeutic frontrunner has yet to emerge.

Experimental background

Among a variety of aggressive and deadly brain tumors we could highlight the glioblastoma. GBM is the most common and malignant brain cancer. Usually, treatment relies on chemotherapy, radiation and surgery. However these treatments are ineffective and the median survival time of a patient is no longer than 15 months (4 to 5 months without health care), due to multifocality of the disease, infiltrative growth and substantial tumor genotypic variability, among other factors [6, 7]. So, nowadays there are no known medical or surgical approaches that constitute

*Correspondence: victor.lopezd@udg.edu
ICREA/Complex Systems Laboratory, Departament de Física, Universitat de Girona, 17071 Girona, Catalonia, Spain

an effective treatment of GBM, and for this reason it is widely considered that the treatment of GBM is likely to benefit from oncolytic virus therapy.

Oncolytic viruses—including retroviruses, herpesviruses and adenoviruses—are an emerging therapy tool for cancers that currently lack effective treatment [8]. The efficiency of different viruses against various tumor cell lines have been studied with in vitro and in vivo experiments [9, 10]. Of these, vesicular stomatitis virus (VSV) has been shown in laboratory studies to have excellent capabilities to become one of the most valuable candidates for virotherapy, due to its very fast lytic cycle and its rapid oncolytic action. In addition, VSV is an enveloped, negative-strand RNA rhabdovirus that can infect a wide variety of species including mice and humans, though it is usually asymptomatic for human beings [11]. Therefore, the anticancer activity of mouse models can be transferable to human trials [12]. This fact makes VSV a strong oncolytic candidate and it has been used in preclinical studies of numerous cancer types, like glioblastomas.

Hence, we focus our attention on the development of a mathematical model of the VSV-GBM virus-tumor system. In the absence of in vivo data, all of the parameter values that we will introduce in the model are extracted from in vitro VSV-GBM experiments. Our main objective is to develop a simple model that can reproduce the VSV-GBM dynamics and explain satisfactorily the experimental in vitro propagation speeds.

Previous mathematical approaches

The most basic mathematical model of the competition between populations was constructed by Alfred J. Lotka and Vito Volterra in 1925 and 1926 independently [13]. For years their model was improved and adapted to different parasite-host systems, including virus infections [14–17]. Nevertheless, we are interested in a specific model which studies the dynamics of an oncolytic virus through a tumor cell population.

In Ref. [5], Wodarz et al. noted that the few previous reaction-diffusion models of oncolytic virus spread [18, 19] include, in addition to basic spatial dynamics, one or more additional assumptions that introduce further complexity. In contrast, they opt for a very simple approach to the infection process with spatial dynamics. The process of adsorption of a virus V by a susceptible tumoral cell T (with rate k_1), and replication of Y viruses that leave each infected cell I (with rate k_2), is essentially described by the reactions

$$V + T \xrightarrow{k_1} I \xrightarrow{k_2} Y \cdot V. \tag{1}$$

Wodarz et al. study the behavior of an in vitro adenovirus in human embryonic kidney epithelial cells, experimentally and computationally, developing a simple model with two equations (see Eqs. (5) and (6) below), one for susceptible tumoral cells and one for infected cells. They make use of partial differential equations (PDEs) to model the virus-tumor system, because PDEs provide efficient information on the spatial and reactive mechanisms affecting the wave propagating fronts and PDEs can be used to compute their speeds.

The model by Wodarz et al. [5] is a two-equation system that was derived from a three-equation model due to Nowak and May [20]. Including diffusion and logistic growth, the Nowak-May model is

$$\frac{\partial [V](r,t)}{\partial t} = D_V \frac{\partial^2 [V](r,t)}{\partial r^2} + k_2 Y [I](r,t) - k_3 [V](r,t), \tag{2}$$

$$\frac{\partial [T](r,t)}{\partial t} = D_T \frac{\partial^2 [T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[I](r,t)+[T](r,t)}{k}\right\} - k_1 [V](r,t)[T](r,t), \tag{3}$$

$$\frac{\partial [I](r,t)}{\partial t} = D_I \frac{\partial^2 [I](r,t)}{\partial r^2} - k_2 [I](r,t) + k_1 [V](r,t)[T](r,t), \tag{4}$$

where $[T]$, $[I]$ and $[V]$ are the concentrations of susceptible tumoral cells, infected tumoral cells and viruses, respectively; D_T, D_I and D_V are their diffusion coefficients, a the tumor growth rate, k its carrying capacity, k_3 the decay rate of free viruses, t the time and r the radial coordinate (assuming radial symmetry, as explained in detail below). Some authors [20] have argued that, in some situations, it may be assumed that $\frac{\partial [V]}{\partial t} = 0$ and therefore, in homogeneous systems $\left(\frac{\partial^2 [V]}{\partial r^2} = 0\right)$, Eq. (2) implies that $[V](r,t) = \frac{k_2 Y}{k_3}[I](r,t)$. However, this assumption (free virus in steady-state) could only be applied if the decay rate of virus k_3 is much larger than the decay rate of the infected cell population k_2 [20]. From these arguments, they obtain the two-equation system used by Wodarz et al. [5], namely

$$\frac{\partial [T](r,t)}{\partial t} = D_T \frac{\partial^2 [T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[I](r,t)+[T](r,t)}{k}\right\} - b[I](r,t)[T](r,t), \tag{5}$$

$$\frac{\partial [I]\,(r,t)}{\partial t} = D_I \frac{\partial^2 [I]\,(r,t)}{\partial r^2} - k_2 [I]\,(r,t) + b[I]\,(r,t)[T]\,(r,t), \tag{6}$$

where $b = \frac{k_1 k_2 Y}{k_3}$.

However, we find two drawbacks in the model (5)–(6) to explain our VSV-GBM system. First, Wodarz assumes $\frac{\partial [V]}{\partial t} = 0$, and thus $[V] \propto [I]$. As said before, this may be valid when $k_3 \gg k_2$ and in some non-spatial models [20] but this is in general not valid for the spatial propagation of virus infections. In such cases, at points located far away from the initially infected area, before the arrival of the infection front we have $[V] = 0$, when the infection arrives $[V] \neq 0$, and after all viruses (and infected cells) have decayed, we have again $[V] = 0$. Therefore, when dealing with spatial infection fronts we have $\frac{\partial [V]}{\partial t} = 0$ only at early and late times, but $\frac{\partial [V]}{\partial t} > 0$ when the first viruses arrive and $\frac{\partial [V]}{\partial t} < 0$ after the passage of the infected front. Moreover, our experimental data (see "Parameter values" section) suggest that in our system k_3 is very close to k_2 and therefore, the assumption $k_3 \gg k_2$is not satisfied here either. Therefore, in contrast to Ref. [5], we cannot assume $\frac{\partial [V]}{\partial t} = 0$, thus we deal with three differential equations (for viruses, susceptible tumoral cells, and infected tumoral cells).

Our second objection to the model (3)–(2) [and its simplification (5)–(6)] is that, according to the first reaction in Eq. (1), virus adsorption causes not only the same decrease in susceptible tumor cells [last term in Eq. (3)] as the increase in infected cells [last term in Eq. (4)], but also the same decrease in viruses. Thus a term $-k_1 [V]\,(r,t)[T]\,(r,t)$ is missing in the right side of Eq. (2), in agreement with many previous works on virus infections [15–17, 21].

In the next section we develop a model which takes both points into account, as well as other important effects (namely, time-delay effects).

Methods

Mathematical models

Here we want to develop a simple, but complete model to understand the dynamics of a virus-tumor system. The theoretical model should be able to explain an in vitro experiment where a virus injected into the center of a tumor spreads through the tumor cell population in a basically two-dimensional geometry. Therefore, we can think of the virus-tumor system as formed by two fronts of propagation, which could be represented as two concentric circles if we assume radial symmetry. The diagram in Fig. 1 illustrates this idea. The outer circle represents the tumor cells, which spread to the outside through a non-specific medium. The inner circle represents the viruses spreading within the tumor. Viruses diffuse through the medium before infecting tumor cells. When infected cells die, a new generation of viruses is created and the process begins anew.

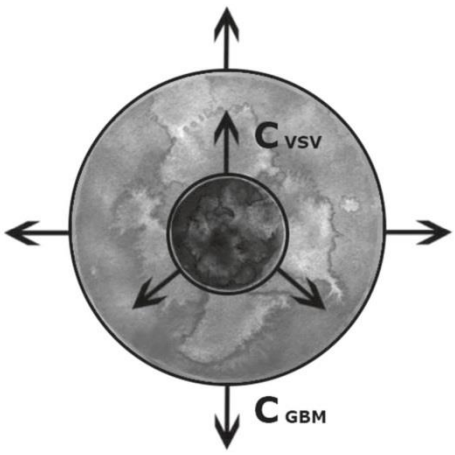


Fig. 1 Two circles representing the two propagation fronts of VSV and GBM. A front of tumor cells spreads radially (*outer circle*). After some time, viruses are inoculated at the center, and a virus front spreads (*inner circle*). If the inner circle expands faster than the outer one ($c_{VSV} > c_{GBM}$), the viruses will eliminate the tumor

The main idea and experimental laboratory data come from Ref. [9], where Wollmann et al. compare nine types of viruses with strong oncolytic potential and conclude that four of them, VSV included, would be worthy of more rigorous studies. Because in subsequent papers [11, 22] they worked with VSV and its recombinant variants or strains, we decided to focus solely on VSV and use these data as experimental basis.

Below we present three increasingly complete (and complicated) models.

Model 1

As a first approach, we adapt the model by Wodarz et al. [5] to the conditions in our VSV-GBM systems, i.e., we do not assume $\frac{dV}{dt} = 0$, and therefore $[V]$ is not proportional to $[I]$ and we need to include the virus dynamics explicitly in the model.

Now the evolution of the virus-tumor system is described by

$$\frac{\partial [V]\,(r,t)}{\partial t} = D_{VSV} \frac{\partial^2 [V]\,(r,t)}{\partial r^2} + F(r,t), \tag{7}$$

$$\frac{\partial [T](r,t)}{\partial t} = D_{GBM}\frac{\partial^2 [T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[I](r,t)+[T](r,t)}{k}\right\} - k_1[V](r,t)[T](r,t), \tag{8}$$

$$\frac{\partial [I](r,t)}{\partial t} = k_1[V](r,t)[T](r,t) - k_2[I](r,t). \tag{9}$$

The first equation describes the evolution of the virus population over time. The viruses can spread ruled by the diffusion coefficient D_{VSV} and the Laplacian (or second space derivative). The function $F(r,t)$ in Eq. (7) incorporates all processes of infection, replication and death and is defined by

$$F(r,t) = -k_1[V](r,t)[T](r,t) + k_2 Y[I](r,t) - k_3[V](r,t). \tag{10}$$

Note that the first term was not included in the models by Nowak-May and Wodarz [Eq. (2)] (see our second objection in "Previous mathematical approaches" section).

Equation (8) describes the change in the number of tumor cells over time. Similarly to viruses, glioblastoma cells can also move, characterized by their own diffusion coefficient D_{GBM}.

Finally, Eq. (9) represents the evolution of infected tumoral cells. We assume that these cells do not move, in agreement Fig. 3D of Ref. [9], where the experiment shows how the infected cells (U-87 MG glioblastoma cells) initially introduced do not move through the host layer throughout the observation period.

Model 2

As we shall see in "Results and discussion" section, model 1 needs further improvements. In model 2 we take into account that infected tumoral cells do not die instantaneously, instead there is a time delay before the cell dies and releases the new progeny of viruses. We will denote this delay or eclipse time as τ and include it into the terms related to the death of infected cells. Thus infected cells will not die proportionally to the density of infected cells at the present time, $k_2[I](r,t)$, but proportionally to the density of infected cells at a previous instant $t-\tau$, $k_2[I](r,t-\tau)$, to properly include this time delay effect on the decay process. It has been shown that the term $-k_2[I](r,t-\tau)$ agrees well with experimental data in a different context (infections of non-tumor cells) [23]. Other reaction-diffusion models do also apply $t-\tau$, although in an alternative way [24, 25]. The differences between their approach and ours is analyzed in Ref. [23].

Therefore, when introducing the delay in the death of infected cells, Eqs. (9) and (10) are modified directly and Eq. (7) changes because the function $F(r,t)$, Eq. (14), is also modified. We do not modify the growth term in Eq. (8) because the reproduction of tumoral cells depends on the total number of tumor cells (infected and susceptible) at that precise instant t. So, we consider the model

$$\frac{\partial [V](r,t)}{\partial t} = D_{VSV}\frac{\partial^2 [V](r,t)}{\partial r^2} + F(r,t), \tag{11}$$

$$\frac{\partial [T](r,t)}{\partial t} = D_{GBM}\frac{\partial^2 [T](r,t)}{\partial r^2} + a[T](r,t)\left\{1 - \frac{[I](r,t)+[T](r,t)}{k}\right\} - k_1[V](r,t)[T](r,t), \tag{12}$$

$$\frac{\partial [I](r,t)}{\partial t} = k_1[V](r,t)[T](r,t) - k_2[I](r,t-\tau), \tag{13}$$

where now

$$F(r,t) = -k_1[V](r,t)[T](r,t) + k_2 Y[I](r,t-\tau) - k_3[V](r,t). \tag{14}$$

This second model is, actually, an approximation of our next model (see Model 3 below).

Model 3

Model 2 takes into account a delay time in the reactive process $I \rightarrow Y \cdot V$, but here we shall see that the delay time also has a very important diffusive effect. The diffusion dynamics of the virus concentration in Eq. (11) is Fickian, which means that it does not take into account the effect of the time delay τ. In year 2002 it was shown [26] that it is very important to take into account that τ is the time interval during which a virus does not move in space (because it is inside an infected cell), thus the delay time should affect the model by slowing down the spread of viruses. Therefore it is necessary to incorporate also this effect to reach a realistic model. For this reason, Eq. (11) must be replaced by an equation with second-order terms to include this diffusive time-delay effect [17, 26, 27].

Thus, finally we describe the spatial-time dynamics of the whole system with the following equations:

$$\frac{\partial [V](r,t)}{\partial t} + \frac{\tau}{2}\frac{\partial^2 [V](r,t)}{\partial t^2} = D_{VSV}\frac{\partial^2 [V](r,t)}{\partial r^2} + F(r,t) + \frac{\tau}{2}\left.\frac{\partial F(r,t)}{\partial t}\right|_g, \tag{15}$$

$$\frac{\partial [T]\,(r,t)}{\partial t} = D_{GBM}\frac{\partial^2 [T]\,(r,t)}{\partial r^2} + a[T]\,(r,t)\left\{1 - \frac{[I]\,(r,t)+[T]\,(r,t)}{k}\right\} - k_1[V]\,(r,t)[T]\,(r,t), \tag{16}$$

$$\frac{\partial [I]\,(r,t)}{\partial t} = k_1[V]\,(r,t)[T]\,(r,t) - k_2[I]\,(r,t-\tau), \tag{17}$$

where the terms proportional to τ in Eq. (15) are the new, second-order terms. A self-contained derivation of Eq. (15) can be found in Ref. [23], Appendix A.

In Eq. (15) $F(r,t)$ is again given by Eq. (14), and Eqs. (12) and (13) from model 2 remain unchanged (Eqs. (16) and (17), respectively).

Note that $F(r,t)$ can be understood as the variation of $[V]$ over time due to all reactive processes, but not to diffusive processes, i.e. $F(r,t) = \left.\frac{\partial [V](r,t)}{\partial t}\right|_g$. This allows the proper calculation of the first time derivative as [17, 27]

$$\left.\frac{\partial F(r,t)}{\partial t}\right|_g = -k_1 F(r,t)[T]\,(r,t) - k_1[V]\,(r,t)\frac{\partial [T]\,(r,t)}{\partial t} + k_2 Y\frac{\partial [I]\,(r,t-\tau)}{\partial t} - k_3 F(r,t). \tag{18}$$

For systems in which the infected cells diffuse appreciably (not our case, see the last paragraph in the model 1 section), an age-structure model with this additional diffusive-delay effect has been proposed by Gourley and Kuang in Ref. [24], p. 558.

In the equation describing the virus dynamics, Eq. (15), we include corrections only up to second order [17, 27]. It has been shown in previous work [26] that the divergence between second-order approximation and full time-delayed equations is small, and thus we can exclude terms of higher orders.

Front speeds

Virus front

Using models 1–3 above, we look for realistic travelling-wave speeds for both the propagation front of viruses (inner front, Fig. 1) and the propagation front of tumor cells (outer front, Fig. 1). Finding the propagation speeds will allow us to compare to the in vitro experiments in order to validate our approach.

In all models 1–3, we can transform the problem into a single-variable system by using the co-moving coordinate $z = r - ct$. Like in previous works [15, 26], we assume the concentration of the three populations at the leading edge of the moving front ($z \longrightarrow \infty$) can be written as $[T] = k - \epsilon_T \cdot \exp(-\lambda z)$, $[I] = \epsilon_I \cdot \exp(-\lambda z)$ and $[V] = \epsilon_V \cdot \exp(-\lambda z)$, thus we assume that tumoral cells are nearly at maximum concentration at large distances from the inoculation point of the viruses, while viruses and infected cells are barely present. We make use of this transformation because beyond the edge of the front of infected cells and viruses, there is only a continuous medium of tumor cells. For non-trivial solutions to exist, the determinant of the matrix corresponding to the linearized model must be zero. The characteristic equations for models 1, 2 and 3 are, respectively,

$$(\lambda c + k_2)\left(\lambda c - D_{VSV}\lambda^2 + kk_1 + k_3\right) - kk_1k_2Y = 0, \tag{19}$$

$$\left(\lambda c + k_2 e^{-\lambda c\tau}\right)\left(\lambda c - D_{VSV}\lambda^2 + kk_1 + k_3\right) - kk_1k_2Ye^{-\lambda c\tau} = 0, \tag{20}$$

$$\left(\lambda c + k_2 e^{-\lambda c\tau}\right)\left[\lambda c - D_{VSV}\lambda^2 + kk_1 + k_3 + \frac{\tau}{2}\left(\lambda^2 c^2 - k^2k_1^2 - 2kk_1k_3 - k_3^2\right)\right] - kk_1k_2Ye^{-\lambda c\tau}\left[1 + \frac{\tau}{2}(\lambda c - kk_1 - k_3)\right] = 0. \tag{21}$$

According to marginal stability analysis [28], the propagation front moves with the minimum possible speed. Therefore,

$$c_{VSV} = \min_{\lambda>0}\left[c\,(\lambda)\right], \tag{22}$$

where $c\,(\lambda)$ is given implicitly by Eqs. (19), (20) and (21). From Eq. (22) we can numerically estimate the speed of VSV infection.

The resulting propagation speeds for models 1–3 will be calculated and plotted in "Results and discussion" section.

We also solve the third model by numerical integration and find the front speed from the position of the virus front wave in a successive steps of time.

Glioblastoma front

Under the hypothesis of two propagation fronts, as shown in Fig. 1, the outermost front would correspond the tumor cells, $[T]$ (GBM in our case of study). In the conditions near this front, all models can be greatly simplified since here the populations of viruses and infected cells are zero (see the outer circle in Fig. 1 for a better understanding), so $[V]\,(r,t) = 0$ and $[I]\,(r,t) = 0$. Hence, it is only necessary to work with the equation for the tumoral cells, Eq. (16) for example, but remembering that $[V]\,(r,t) = [I]\,(r,t) = 0$,

$$\frac{\partial [T]\,(r,t)}{\partial t} = D_{GBM}\frac{\partial^2 [T]\,(r,t)}{\partial r^2} + a[T]\,(r,t)\left\{1 - \frac{[T]\,(r,t)}{k}\right\}. \tag{23}$$

At the leading edge of this front, we assume that $[T](r,t) = \epsilon_T \cdot \exp(-\lambda z)$, and after some algebra we easily obtain the speed of the glioblastoma front,

$$c_{GBM} = 2\sqrt{D_{GBM}\, a}, \tag{24}$$

where D_{GBM} is the glioblastoma diffusion coefficient and a the growth rate, both estimated in the next subsection. Note that Eq. (24) is the well-known Fisher propagation speed [29]. Some recent extensions have been proposed [6, 30], but they are not necessary for the purposes of the present paper.

Parameter values

We estimate most of our parameters from in vitro experiments on VSV applied to GBM [9, 11, 22]. The parameters that we could not draw from such experiments have been obtained from other rigorous studies on VSV or glioblastoma.

We use two different values of D_{VSV} because the diffusion coefficient of VSV has not been measured in gliomas. The only value of VSV available (measured in an specific water solution) is $D_{VSV} = 8.37 \cdot 10^{-5}$ cm^2/h [31]. Another value measured in agar of VSV-similar viruses is $D_{VSV} = 1.44 \cdot 10^{-4}$ cm^2/h [17].

Concerning D_{GBM}, Stein et al. [32] performed an in vitro experiment in which a GBM tumor spheroid is implanted into a collagen gel. The diffusion coefficient is measured by tracking individual cells on the first day, calculating their motion and averaging over many cells. Stein and coworkers measure diffusion coefficients in the radial and angular directions, which lead to the value $D_{GBM} = 3.75 \cdot 10^{-6}$ cm^2/h [6].

Besides spreading, the number of cells also increases. The parameter a is the corresponding proliferation rate. In vitro measurements provide ample scope for this parameter, $0.04 < a < 0.3$ day^{-1} [33], and similarly in vivo studies yield $0.01 < a < 0.14$ day^{-1} [34].

The saturation cell density, k, measures the maximum concentration of tumor cells (susceptible and infected) per unit volume that the system can support, and its usual value is $k = 10^6$ cells/cm^3 (e.g.,Refs. [35, 36]).

We next analyze the rest of parameters, which are calculated from the experimental studies by Wollmann et al. [9, 11, 22].

The yield or burst size Y represents the total amount of viruses produced by the death of a single infected cell. There is no accurate numerical value calculated for the case of VSV infecting GBM. However, by studying Fig. 4 in Ref. [11] we can obtain an estimation. The burst size can be understood as the ratio between the maximum and initial number of viruses, i.e. $Y = \frac{V_{\max}}{V_0}$. From that figure, V_0 is between $10 - 100$ PFU/ml (last two plots in Fig. 4 in [11]) and $V_{\max}$ between $10^8 - 10^9$ PFU/ml (the maximum is reached between 1 and 2 days post infection), so we conclude that $10^6 < Y < 10^8$. This also agrees with the value measured in Ref. [37], although in that case VSV infects BHK-21 cells (not GBM cells).

We have seen that there is a time lapse between a cell being infected by a virus and that cell dying (and therefore, adding more viruses to the system). This time lapse is called the delay time, τ. It plays a main role in the virus propagation speed, but has not been accurately measured. From the in vitro experiments described in Ref. [9] we can try to estimate the value of τ. On one hand, we know that the death of infected cells begins about 6 hours post infection (hpi) of the virus to susceptible tumoral cells. We also know that infected cells can be seen as early as 4 hpi (they are tracked down using GFP fluorescence). From both data, we conclude that viruses leave infected cells at least 2 h after infection. On the other hand, in a different experiment infected cells are added directly (rather than infecting viruses) and new infected cells were detected after 12 h. This period includes the time needed for the viruses to multiply within the infected cells, leave the cell and infect new tumoral cells. So we can also assume that τ must be lower than 12 h. In summary, we will work with the range $2 < \tau < 12$ h.

The adsorption rate, k_1, describes the efficacy of the whole infection process (rate of virus entry and probability of successful infection). The value of k_1 could be measured in an experiment where the reproduction of viruses and host cells were prevented. Such an experiment has been performed for other viruses [38] but not for VSV infecting GBM. Since we do not have the ideal conditions in the experiments cited before [9, 11, 22], we will use the earliest data post-inoculation available in the experimental data in Ref. [11] to minimize the effect of reproduction and thus obtain the best possible estimation for k_1.

Equations (7) and (8) are simplified in the absence of reproduction and natural death, and when the population is studied as a whole (i.e. ignoring diffusion terms) we have

$$\frac{d[V](t)}{dt} = \frac{d[T](t)}{dt} = -k_1[V](t)[T](t). \tag{25}$$

Obviously, integrating we get $[T](t) = [V](t) + \xi$, where ξ is the constant of integration. Note that ξ is the difference between the concentrations of tumor cells and viruses. In order to estimate k_1, we can rewrite the previous Eq. (25) as $\frac{d[T](t)}{dt} = -k_1[T](t)([T](t) - \xi)$ and making the necessary algebra we obtain the final formula for calculating the adsorption rate,

$$k_1 = \frac{1}{\xi\,(t - t_0)}\left[\ln\left(\frac{T}{T-\xi}\right) - \ln\left(\frac{T_0}{T_0 - \xi}\right)\right]. \tag{26}$$

It is difficult to know the exact concentration of cells at the beginning of the experiment or at certain time t, because only relative concentrations were reported. However, extrapolating data provided in the previous cited

papers by Wollmann et al. (Fig. 3C Control in [11], bar G/GFP), we believe it is correct to assume that the values of initial tumor cells lie in the range $T_0 = 10^6 - 10^8$ cells/cm^3, and that $T = 0.65T_0$ cells/cm^3, $t - t_0 = 36$ h. This allows the calculation of the adsorption rate, as $5 \cdot 10^{-10} < k_1 < 5 \cdot 10^{-8}$ cm^3/h. This is a rather wide range, but we show in "Effect of k_1 and Y" section that k_1 (as well as Y) does not overly affect the propagation front speed of VSV.

Finally, parameters k_2 and k_3 correspond to the rates of death of infected cells and virus, respectively. Therefore, the average life-time of an infected cell and a virus are $1/k_2$ and $1/k_3$, respectively.

The rate of death of infected cells k_2 could be also understood as the growth of viruses. Thus, for $t < \tau$ no new virus are seen in the corresponding experiment (because no infected cell has died yet), but for $t \geq \tau$ the infected cells start to die ruled by $dI = -k_2 I_0 dt$. The death of each infected cell produces Y virus, thus $dV = -YdI = k_2 Y I_0 dt = k_2 V_{\max} dt$. Integrating, we get $k_2 = \frac{V_{\max} - V_0}{\Delta t \cdot V_{\max}} \approx \frac{1}{\Delta t} = \frac{1}{t^* - \tau}$, where t^* represents the time when the virus population reaches its maximum. According to Fig. 4B in Ref. [11], experimental data (labeled as VSV-G/GFP) show that the maximum is reached at $t^* = (48 \pm 12)$ h. Nevertheless, the final result of k_2 will depend on τ and we have a range rather than a single value for τ (see above). Note, however, that for model 1 there is no time delay, so k_2 is calculated straightforwardly as the inverse of time t^* at which the concentration of viruses reaches its maximum, $k_2 = \frac{1}{t^*}$ h^{-1}. Models 2 and 3 are dealt with in "Results and discussion" section.

The evolution of the viruses over time in an environment where they die but cannot reproduce is ruled by $dV = -k_3 V dt$. Through simple integration we get $V(t) = V_0 \exp[-k_3 (t - t_0)]$. In the same experiment as before, Fig. 4B in Ref. [11], we now have two cases where these conditions are exactly reproduced (because VSV-dG-GFP and VSV-dG-RFP are replication-restricted virus variants, so they basically die). We can estimate both values of k_3 from the experimental data, namely $V(t = 24\,\text{h}) = 30$ PFU/cm^3, $V(t = 48\,\text{h}) = 20$ PFU/cm^3 and $V(t = 72\,\text{h}) = 8$ PFU/cm^3 for the mutant dG-GFP and $V(t = 24\,\text{h}) = 12$ PFU/cm^3, $V(t = 48\,\text{h}) = 8$ PFU/cm^3 and $V(t = 72\,\text{h}) = 6$ PFU/cm^3 for dG-RFP. Performing linear fits to $\ln V$ versus t, we obtain that $0.014 < k_3 < 0.028$ h^{-1}.

Results and discussion

GBM and VSV front speeds: theory versus experiment

Our main objective is to obtain realistic values for the propagation speeds in an in vitro virus-tumor system, providing positive results from a biophysical point of view for the realization of these treatments.

In "Methods" section we have described three possible models for our VSV-GBM system and the necessary experimental parameter values. Here we present the speeds predicted by these models.

The case of tumor expansion has a single, simple solution for all models, Eq. (24), since the infection does not play a role here. Substituting the values of D_{GBM} and a we obtain that $c_{GBM} = 2.5 \cdot 10^{-4}$ cm/h, with $a = 0.1$ day^{-1}, which we think is a reasonable mean value. Indeed, the range of measurements of the proliferation rate is $0.01 < a < 0.3$ day^{-1}, which yields a range of speeds $7.9 \cdot 10^{-5} < c_{GBM} < 4.33 \cdot 10^{-4}$ cm/h). Stein and coworkers measured an experimental in vitro speed range of $2.37 \cdot 10^{-4} < c_{GBM} < 5.54 \cdot 10^{-4}$ cm/h [33], which is consistent with our model, despite the simplicity of Eq. (24).

The case of the virus front is less straightforward. As we have already discussed in "Parameter values" section, a very important but not strictly well-measured parameter is the delay time τ. Therefore, the speed results have been calculated in terms of this parameter, $c(\tau)$. The death rate of infected cells k_2 also changes, because it depends directly on τ.

The infection front speed, c_{VSV}, can be seen in Fig. 2. For each of the 3 models we have plotted the results from typical parameter values (bold lines). To compute these results we have chosen the parameter values that seem to be the most representative and accepted for this experiment: average values of k_2 and k_3, the value of D_{VSV} calculated for VSV in an specific water solution and the larger values of k_1 and Y. However we have also computed c_{VSV} by varying each of the parameters of Eqs. (19)–(21), with the exception of k because $k = 10^6$ cells/cm^3 is a widely accepted value in research papers (see "Parameter values" section). In Fig. 2 we include the upper and lower bounds for the front speed obtained, for each of the 3 models, from the experimental parameter ranges (parameter values are specified at the caption).

The hatched area in Fig. 2 corresponds to the experimental values of VSV speed estimated from the in vitro experiment by Wollmann et al. in Ref. [9], Fig. 3A.

Dotted lines correspond to the analytical results to model 1, Eqs. (7)–(10), i.e. the classical model adapted from the equations in Ref. [5]. Obviously they are horizontal lines, since they do not depend on τ. As we can see in Fig. 2, model 1 yields speeds much faster than the experimental observations. The curves are the numerical results from our time-delayed reaction-diffusion models. Dashed curves correspond to model 2, given by Eqs. (11)–(14). We see that just by taking into account the eclipse or delay time on the death of infected cells, we obtain much better results as compared with experimental velocities, although not enough to satisfactorily explain the data (the

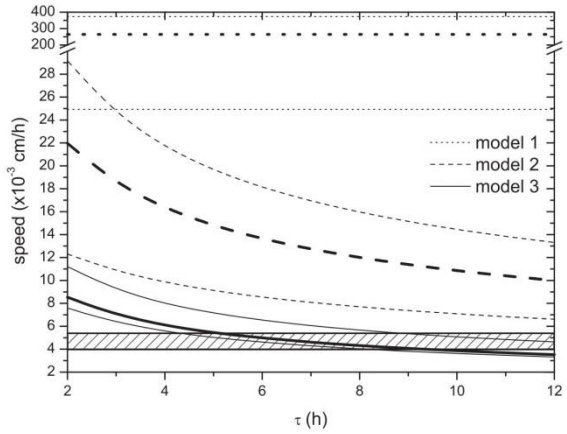


Fig. 2 VSV front propagation speed as a function of the delay time τ, for model 1 (dotted lines), model 2 (dashed curves) and model 3 (solid curves). The hatched area shows the experimental in vitro VSV front speed [9]. Upper bounds are computed from: $k_1 = 5 \cdot 10^{-8}$ cm^3/h, $k_2 = \frac{1}{36-\tau}$ h^{-1} ($k_2 = \frac{1}{36}$ h^{-1} for model 1), $k_3 = 0.014$ h^{-1}, $Y = 10^8$ and $D_{VSV} = 1.44 \cdot 10^{-4}$ cm^2/h. Lower bounds are computed from: $k_1 = 5 \cdot 10^{-10}$ cm^3/h, $k_2 = \frac{1}{60-\tau}$ h^{-1} ($k_2 = \frac{1}{60}$ h^{-1} for model 1), $k_3 = 0.028$ h^{-1}, $Y = 10^6$ and $D_{VSV} = 8.37 \cdot 10^{-5}$ cm^2/h. The results from typical values (bold lines) are computed from: $k_1 = 5 \cdot 10^{-8}$ cm^3/h, $k_2 = \frac{1}{48-\tau}$ h^{-1} ($k_2 = \frac{1}{48}$ h^{-1} for model 1), $k_3 = 0.02$ h^{-1}, $Y = 10^8$ and $D_{VSV} = 8.37 \cdot 10^{-5}$ cm^2/h. In all the cases $k = 10^6$ cells/cm^3

minimum bound of model 2 in Fig. 2 is above the hatched area). Finally, solid curves in Fig. 2 correspond to model 3 (please recall that this is extremely close to the full time-delayed equation, see "Methods" section). The equations for this main model, Eqs. (15)–(18), when considering typical parameter values, produce results that agree with the experimental data within a range of τ between 5.0 and 9.3 h.

According to our best description (model 3), the entire range of speed c_{VSV} in Fig. 2 is an order of magnitude faster than the speed of propagation of glioblastoma, c_{GBM}, (see above). Therefore the virus front could theoretically reach the tumoral front and infect it all. We stress that this is a model appropriate for in vitro experiments, whereas in vivo more complex models will be necessary (as discussed below).

In Fig. 3 we show snapshots of the viruses and infected cells profiles as functions of the radial axis, computed from the computational simulations at three time instants. The simulations have been performed by numerical integration of model 3, which is biologically more realistic and produces results in agreement with the experimental data (see Fig. 2). We use the typical parameter values used in Fig. 2 (bold lines, see caption for the values). We can see in Fig. 3 that both propagation fronts advance at the same speed and with regular shapes.

From the profiles we can see that the number of infected cells grows rapidly, then there is a plateau of infected cells (as a result of the time delay τ before any infected cell dies), and then decay at a rate k_2. The virus profiles show an abrupt rise when infected cells start dying (end of the plateau of infected cells) and then keep rising up to a peak. Behind this peak, the virus death term k_3 predominates over the virus production, and the number of viruses decay. Although Fig. 3 seems to indicate that the front of infected cells appears prior to the virus front, the opposite happens (this can be appreciated by enlarging the vertical scale).

From these simulations we can calculate the front speed by tracking the position of the edge of the front of the virus at successive steps of time. A simple space vs time data is generated and then, the front speed is directly the slope. From the simulations (parameter values are the same than typical values in Fig. 2 with $\tau = 6$ h) we find a front speed of $4.829 \cdot 10^{-3}$ cm/h. The relative error between the simulations and the analytic speed [$c_{VSV} = 4.853 \cdot 10^{-3}$ cm/s, from Eqs. (21)–(22)] is only about 0.5 %.

An alternative way to know the front propagation speed from Fig. 3 is the plateau of infected cells. Its width is directly related with the time delay τ and the infection front speed as $width = \tau \cdot c$. Then, the result for the speed is $(0.53858 - 0.51317)$ cm / 6 h $= 4.735 \cdot 10^{-3}$ cm/h (distances for $t = 108$ h), and the relative error (compared with the analytical results with same parameter values than the simulations) is lower than 2.5 % ($c_{VSV} = 4.853 \cdot 10^{-3}$ cm/s).

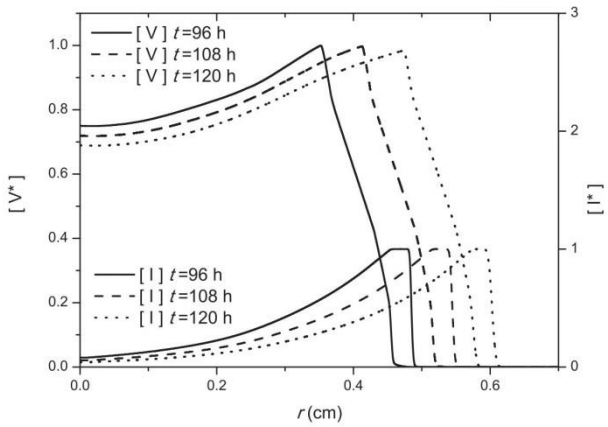


Fig. 3 Radial profiles of $[V^*]$ and $[I^*]$ at three different times for model 3. The labels of V^* and I^* stand for the units used, defined as $\frac{[V]}{[V]_{max}}$ and $\frac{[I]}{[I]_{max}}$, respectively. The profiles are computed from numerical integration

Effect of k_1 and Y

In "Parameter values" section we have estimated the values of the parameters used in our mathematical models. Some of them, e.g. D_{VSV}, D_{GBM} and k, have well-defined values, which are taken from the references indicated in the text. The delay time τ plays a very important role and therefore we have found the front propagation speed as a function of this parameter (remember that $k_2 = \frac{1}{48-\tau}$, so we could add k_2 to this argument). Other parameters like a and k_3 have a range of possible values, albeit a narrow one, and as such we use the mean value, or that usually accepted by other sources. Lastly, parameters Y and k_1 have very wide ranges, spanning several orders of magnitude, but as we shall show below, they do not have an important effect on the virus front speed.

In Fig. 4 the speed of VSV is calculated from model 2 (Eqs. (11)–(14)) and model 3 (Eqs. (15)–(18)). Setting the typical parameter values previously used in Fig. 2 (bold curves) and Fig. 3 for D_{VSV}, D_{GBM}, k, k_3 and the average value $\tau = 8$ h (so $k_2 = 1/40\ \text{h}^{-1}$), which yields results consistent with the range of experimental speeds (Fig. 2), we have varied the values of Y and k_1 for each of both models.

In model 2 (upper curves in Fig. 4) the speed dependence on Y and k_1 is fairly important. Indeed, by increasing these variables by two orders of magnitude, the speed increases on average by 25 and 18 %, respectively. However, looking at the best approach, model 3 (lower curves), we note that the speed increases only by 3 and 2 % for Y and k_1, respectively.

Therefore, model 3 has little dependence on the parameters Y and k_1 and the delay time is the most important parameter (Fig. 2). In contrast, model 2 depends more directly on both parameters, although τ still remains the crucial one (compare the change of the speed in Fig. 2 with those in Fig. 4 for model 2). To obtain a speed of virus propagation similar to the observed data ($c \approx 5 \cdot 10^{-3}$ cm/h) with model 2, we should modify Y and k_1 out of the experimental ranges. Indeed, their values should be about $Y = 10^4$ or $k_1 = 5 \cdot 10^{-12}$ cm^3/h. Therefore, we could get a speed in agreement with the experimental data, but only using unrealistic parameter values, which do not correspond to VSV. This is further proof that our final model 3, the time-delayed reaction-diffusion set of equations, is a good mathematical tool to explain this kind of virus-tumor biological systems.

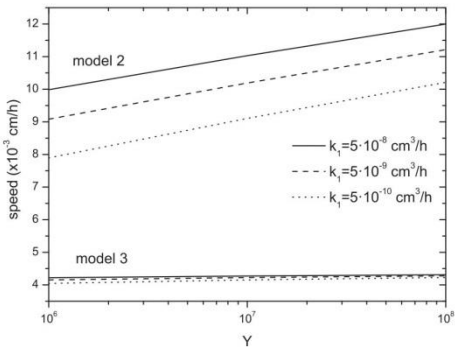


Fig. 4 VSV invasion speed on GBM for various values of Y and k_1. The other parameter values are $k = 10^6$ cm^{-3}, $k_2 = \frac{1}{40}$ h^{-1}, $\tau = 8$ h and $k_3 = 0.02$ h^{-1}. Model 3 proves that neither Y nor k_1 affect much the speed of the front

Conclusions

A simple set of time-delayed equations have been built to understand the dynamics of a virus-tumor system. We have improved a previous model with new ideas and carefully incorporated experimental results (especially Ref. [9]). Figure 2 proofs that our best framework (model 3) is in reasonable agreement with the experimental data. Furthermore, the figure shows that neither model 1 nor model 2 can explain the experimental data. So it is absolutely necessary to add the second-order terms proportional to τ in Eq. (15) to properly include the time-delay effect.

We have shown that the delay time τ is the crucial parameter in our models (even when compared to other parameters that are strongly unknown, such as k_1 and Y). As we could have expected, as τ increases, the speed of the virus front decreases, because viruses spend more time inside the cell, and therefore at rest. In spite of being of utmost importance, the role of the delay or eclipse time has not been taken into account in previous models of virus treatment of tumors [5, 18, 19].

We have found that our new model can satisfactorily predict the front speed for the lytic action of oncolytic VSV on glioblastoma observed in vitro. But this is only a first step towards a deep biophysical understanding of the principles of virus-tumor space-time spread in a complex system. This model could be extended to be applied to in vivo experiments where, among other effects, the immune response should be also included in the model because it may play a significant role regulating the efficacy of the therapy. In particular, it seems that there is currently no agreement about which approach is better in oncolytic therapy, whether to modify oncolytic viruses to obtain the maximum antitumoral immune response [39], to transiently suppress the immune response [40], or to use a combination of both [40]; future appropriate modeling of the three scenarios might help in tackling this controversy from a different perspective.

In this paper we have focused on GBMs because of the experimental data available, but our model could apply also to many non-diffusive cancers, for which viral therapy is a promising approach [18, 19, 41], since the reaction-diffusion equations for the viruses [Eqs. (15)–(18)] will still be valid, even though in such cases tumor cells will not diffuse. Thus, we provide a basis that can be applied in the near future to realistically simulate in vivo virus treatments of several cancers.

Reviewers' comments

Reviewer's report 1

Yang Kuang, Arizona State University, United States of America

Reviewer comments: The paper is mostly well written with only a few places where I can suggest the authors to consider adding more details or be aware of alternative explanations.

1: The authors made a valid point that $\frac{\partial [V]}{\partial t}$ is not always close to 0. However, a routine argument used in the mathematical modeling community is the quasi-state-steady approximation. This argument suggests that due to virus' fast dynamics (virus reproduces probably in less than one hour once the first virus reproduced), over the longer period tumor cell growth time (of days), on average, the total virus amount changes at a rate far less than the maximum possible rate when all viruses reproduce at the maximum rate. Mathematically, one can show this quasi-steady-state level can be approximated by setting $\frac{\partial [V]}{\partial t} = 0$ and solving V in terms of other variables. 2: A better reference in the virus modeling context for the need of adding the virus loss term $-k_1 [V](r,t)[T](r,t)$ may be E. Beretta and Y. Kuang: Modeling and analysis of a marine bacteriophage infection. Math. Biosc. 149, 5776(1998), where each and every term is carefully explained in the context of biology. 3: The justification for the Eq. (15) is mathematically simple, but mechanistically very ad hoc and difficult to follow. A possible alternative way to modeling the delay dependence of the diffusive action is to assign virus an age. A good reference on this approach is S. A. Gourley and Y. Kuang: A Delay Reaction-Diffusion Model of the Spread of Bacteriophage Infection, SIAM J. Appl. Math., 65, 50566(2005). 4: I think readers will benefit if the authors can provide more about the data nature and even a figure which may suggest that the VSV front is as described in Fig. 2. The authors may take a look of our recent work on in vitro GBM modeling and wave speed estimation to see how we handled this. Tracy L. Stepien, Erica M. Rutter, and Yang Kuang, 2015. A data-motivated density-dependent diffusion model of in vitro glioblastoma growth, Math. Biosc. Eng., 12, 11571172.

Authors' response: *We want to thank Dr. Y. Kuang for his revision of our manuscript and the suggestions provided to make it more complete and comprehensive. We answer each of his four comments separately below:*

1. The quasi-state-steady approximation is truly widely used in mathematical modeling. It implies that the virus dies in a very short time, and the rate of the virus producing infected cells is short enough not to create a great amount of viruses. Mathematically, this means that, $k_3 \gg k_2$ [20]. This condition is not fulfilled in our VSV-GBM system, where $k_3 \approx k_2$. As a result, we consider that, in such a system, it is better to develop our model and perform the calculation with all three equations. We explain this before Eqs. (5) and (6).

2. We have added the relevant reference suggested at the end of "Previous mathematical approaches" section.

3. The justification of Eq. (15) is described in more detail in Ref. [23], *Appendix A. We mention this below Eq. (17).*

We also cite [below Eq. (18)] the interesting reference suggested, which applies to systems in which infected cells diffuse (not our case).

4. A new figure (Fig. 3) has been added to the paper showing the evolution in space and time of the concentration of virus and infected cell populations. We have computed these profiles through numerical integration and they now provide a new source from where to calculate the front speed for model 3. We see that this new value agrees with the experimental data found in Wollmann et al. experiments and with our analytical results. Readers will probably benefit from this new approach in order to completely understand the significance of our new equations and the good agreement between theoretical and experimental data. So, we specially appreciate the advice (from both referees) to include this kind of results.

Reviewer's report 2

Georg Luebeck, Fred Hutchinson Cancer Research Center, United States of America

Reviewer comments: The mathematical framework presented by de Rioja et al. for oncolytic infection of GBM cells by the VSV virus and its impact on tumor growth in culture builds upon previous modeling. The authors show, convincingly (at least for the infection experiments in GBM cells), that it is important to include a time delay that represents the time from infection to cell death and production of new viral particles. Furthermore, the case is made that the time delay effect and sequestration of the virus in the infected cells leads to second order effects which further slow the spread of the virus.

Although the model is rather simplistic (it has radial symmetry, no vasculature, infection starting from a single point) and most kinetic rates are only known imprecisely, the agreement of the model prediction with the experimental data on the front speed of the VSV action is reassuring that the mathematical description of the augmented model is biologically plausible. The conclusions, of course, would have been stronger had the authors used an independent experimental model to validate their finding. Also, there is no notion of uncertainty in the predictions shown in Fig. 2. It would be useful if a sensitivity analysis could be included to demonstrate that model 3 is indeed the only model (among the 3) that is consistent with the experimental data.

Also, it is somewhat surprising that the authors did not also visualize the solutions of their models as radial density 'snapshots' at various endpoints. This (together with the parameter values used) could help others to reproduce their results.

Authors' response: *We thank the review by Dr. G. Luebeck, which is quite positive with our research manuscript. We have reviewed the text according to his suggestions.*

As suggested, we have improved Fig. 2 by adding lower and upper bounds to the model predictions obtained by considering the whole range of the parameter uncertainties. The new figure shows how important it is to include the second-order terms, because neither model 1 nor model 2 can explain the experimental data. The robustness of this conclusion has improved with this sensitivity analysis.

A new Fig. 3 has been added to the manuscript following the advice of both referees. It shows three snapshots of the populations of virus and infected cells in space at different instant. This new figure provides a visualization of the expansion process, as well as a new way to compute the front speed for model 3.

Minor points have been take into account and corrected in the text.

Abbreviations

GBM: Glioblastoma; VSV: vesicular stomatitis virus; PDE: partial differential equation; PFU: plaque-forming unit; BHK: baby hamster kidney; HPI: hours post infection; GFP: green fluorescent protein.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VLR: participated in the design of the model, did the modeling, the analytic calculations and simulations and drafted the manuscript. NI: supervised the application of the analytical model and parameter selection, did the computational simulations and revised and edited the manuscript. JF: designed the mathematical model, supervised the implementation and participated in writing the manuscript. All authors read and approved the manuscript.

Acknowledgments

This work was partially funded by ICREA (Academia award) and the MINECO (projects SimulPast-CSD2010-00034, FIS-2009-13050 and FIS-2012-31307).

Received: 23 September 2015 Accepted: 11 December 2015
Published online: 07 January 2016

References

1. Kelly E, Russell SJ. History of oncolytic viruses: genesis to genetic engineering. Mol Ther. 2007;15(4):651–9.
2. Ledford H. Cancer-fighting viruses near market. Nature. 2015;526(7575):622–23.
3. Novozhilov AS, Berezovskaya FS, Koonin EV, Karev GP. Mathematical modeling of tumor therapy with oncolytic viruses: Regimes with complete tumor elimination within the framework of deterministic models. Biol Direct. 2006;1:6. doi:10.1186/1745-6150-1-6.
4. Karev GP, Novozhilov AS, Koonin EV. Mathematical modeling of tumor therapy with oncolytic viruses: effects of parametric heterogeneity on cell dynamics. Biol Direct. 2006;1:30. doi:10.1186/1745-6150-1-30.
5. Wodarz D, Hofacre A, Lau JW, Sun Z, Fan H, Komarova NL. Complex spatial dynamics of oncolytic viruses in vitro: mathematical and experimental approaches. PLoS Comp Biol. 2012;8(6):e1002547. doi:10.1371/journal.pcbi.1002547.
6. Fort J, Solé RV. Accelerated tumor invasion under non-isotropic cell dispersal in glioblastomas. New J Phys. 2013;15:055001–10.
7. Özduman K, Wollmann G, Piepmeier JM, van den Pol AN. Systemic vesicular stomatitis virus selectively destroys multifocal glioma and metastatic carcinoma in brain. J Neurosci. 2008;28(8):1882–93. doi:10.1523/JNEUROSCI.4905-07.2008.
8. Wollmann G, Ozduman K, van den Pol AN. Oncolytic virus therapy for glioblastoma multiforme: concepts and candidates. Cancer J. 2012;18(1):69–81. doi:10.1097/PPO.0b013e31824671c9.

9. Wollmann G, Tattersall P, van den Pol AN. Targeting human glioblastoma cells: comparison of nine viruses with oncolytic potential. J Virol. 2005;79(10):6005–22.
10. Freeman AI, Zakay-Rones Z, Gomori JM, Linetsky E, Rasooly L, Greenbaum E, et al. Phase I/II trial of intravenous NDV-HUJ oncolytic virus in recurrent glioblastoma multiforme. Mol Ther. 2006;13(1):221–8.
11. Wollmann G, Rogulin V, Simon I, Rose JK, van den Pol AN. Some attenuated variants of vesicular stomatitis virus show enhanced oncolytic activity against human glioblastoma cells relative to normal brain cells. J Virol. 2010;84(3):1563–73. doi:10.1128/JVI.02040-09.
12. Russell SJ, Peng K-W, Bell JC. Oncolytic virotherapy. Nat Biotechnol. 2012;30(7):658–70. doi:10.1038/nbt.2287.
13. Brauer F, Castillo-Chavez C. Mathematical models in population biology and epidemiology. New York: Springer. 2001123–125.
14. Koch AL. The growth of viral plaques during the enlargement phase. J Theor Biol. 1964;6(03):413–431.
15. Yin J, McCaskill JS. Replication of viruses in a growing plaque: a reaction-diffusion model. Biophys J. 1992;61(6):1540–1549. doi:10.1016/S0006-3495(92)81958-6.
16. Haseltine EL, Lam V, Yin J, Rawlings JB. Image-guided modeling of virus growth and spread. Bull Math Biol. 2008;70(6):1730–48. doi:10.1007/s11538-008-9316-3.
17. Amor DR, Fort J. Virus infection speeds: Theory versus experiment. Phys Rev E. 2010;82:061905. doi:10.1103/PhysRevE.82.061905.
18. Wein LM, Wu JT, Kirn DH. Validation and analysis of a mathematical model of a replication-competent oncolytic virus for cancer treatment: implications for virus design and delivery. Cancer Res. 2003;63(6):1317–24.
19. Mok W, Stylianopoulos T, Boucher Y, Jain RK. Mathematical modeling of herpes simplex virus distribution in solid tumors: implications for cancer gene therapy. Clin Cancer Res. 2009;15(7):2352–60. doi:10.1158/1078-0432.CCR-08-2082.
20. Nowak M, May RM. Virus dynamics: Mathematical principles of Immunology and Virology. Oxford: Oxford University Press; 2000, pp. 100–109.
21. Beretta E, Kuang Y. Modeling and analysis of a marine bacteriophage infection. Math Bioscienc. 1998;149:57–76.
22. Wollmann G, Robek MD, van den Pol AN. Variable deficiencies in the interferon response enhance susceptibility to vesicular stomatitis virus oncolytic actions in glioblastoma cells but not in normal human glial cells. J Virol. 2007;81(3):1479–91.
23. de Rioja VL, Fort J, Isern N. Front propagation speeds of T7 virus mutants. J Theor Biol. 2015;385:112–118. doi:10.1016/j.jtbi.2015.08.005.
24. Gourley SA, Kuang Y. A delay reaction-diffusion model of the spread of bacteriophage infection. SIAM J Appl Math. 2005;65(2):550–566. doi:10.1137/S0036139903436613.
25. Jones DA, Smith HL, Thieme HR, Röst G. On spread of phage infection of bacteria in a petri dish. SIAM J Appl Math. 2012;72(2):670–688. doi:10.1137/110848360.
26. Fort J, Mendez V. Time-delayed spread of viruses in growing plaques. Phys Rev Lett. 2002;89(17):178101.
27. Isern N, Fort J. Time-delayed reaction-diffusion fronts. Phys Rev E. 2009;80:057103.
28. Ebert U, van Saarloos W. Front propagation into unstable states: universal algebraic convergence towards uniformly translating pulled fronts. Physica D. 2000;146:1–99.
29. Fisher RA. The wave of advance of advantageous genes. Ann Eugenics. 1937;7:353–369.
30. Stepien TL, Rutter EM, Kuang Y. A data-motivated density-dependent diffusion model of in vitro glioblastoma growth. Math Biosci Eng. 2015;12(6):1157–1172. doi:10.3934/mbe.2015.12.1157.
31. Ware BR, Raj T, Flygare WH, Lesnaw JA, Reichmann ME. Molecular Weights of Vesicular Stomatitis Virus and Its Defective Particles by Laser Light-Scattering Spectroscopy. J Virol. 1973;11(1):141–145.
32. Stein AM, Vader DA, Deisboeck TS, Chiocca EA, Sander LM, Weitz DA. Directionality of glioblastoma invasion in a 3D in vitro experiment. arXiv, http://arxiv.org/pdf/q-bio/0610031.pdf. Accessed 30 Jul 2015.
33. Stein AM, Demuth T, Mobley D, Berens M, Sander LM. A mathematical model of glioblastoma tumor spheroid invasion in a three-dimensional in vitro experiment. Biophys J. 2007;92(1):356–65.
34. Rockne R, Rockhill JK, Mrugala M, Spence AM, Kalet I, Hendrickson K, et al. Predicting the efficacy of radiotherapy in individual glioblastoma patients in vivo: a mathematical modeling approach. Phys Med Biol. 2010;55(12):3271–85. doi:10.1088/0031-9155/55/12/001.
35. Friedman A, Tian JP, Fulci G, Chiocca EA, Wang J. Glioma virotherapy: effects of innate immune suppression and increased viral replication capacity. Cancer Research. 2006;66(4):2314–19.
36. Eikenberry SE, Sankar T, Preul MC, Kostelich EJ, Thalhauser CJ, Kuang T. Virtual glioblastoma: growth, migration and treatment in a three-dimensional mathematical model. Cell Prolif. 2009;42(04):511–528. doi:10.1111/j.1365-2184.2009.00613.x.
37. van den Pol AN, Davis JN. Highly attenuated recombinant vesicular stomatitis virus VSV-12'GFP displays immunogenic and oncolytic activity. J Virol. 2013;87(2):1019–34. doi:10.1128/JVI.01106-12.
38. Shishido K, Watarai A, Naito S, Ando T. Action of bleomycin on the bacteriophate T7 infection. J Antibiot (Tokyo). 1975;28(9):676–80.
39. Koks CAE, De Vleeschouwer S, Graf N, Van Gool SW. Immune Suppression during Oncolytic Virotherapy for High-Grade Glioma; Yes or No? J Cancer. 2015;6(3):203–217. doi:10.7150/jca.10640.
40. Russell SJ, Peng K-W, Bell JC. Oncolytic virotherapy. Nature Biotech. 2012;30(7):658–70. doi:10.1038/nbt.2287.
41. Mahoney DJ, Stojdl DF, Laird G. Virus therapy for cancer. Sci Am. 2014;311(5):54–9.

www.nature.com/scientificreports

OPEN

The ancient cline of haplogroup K implies that the Neolithic transition in Europe was mainly demic

Received: 9 February 2017
Accepted: 29 August 2017
Published online: 11 September 2017

Neus Isern[1], Joaquim Fort[1,2] & Víctor L. de Rioja[1]

Using a database with the mitochondrial DNA (mtDNA) of 513 Neolithic individuals, we quantify the space-time variation of the frequency of haplogroup K, previously proposed as a relevant Neolithic marker. We compare these data to simulations, based on a mathematical model in which a Neolithic population spreads from Syria to Anatolia and Europe, possibly interbreeding with Mesolithic individuals (who lack haplogroup K) and/or teaching farming to them. Both the data and the simulations show that the percentage of haplogroup K (%K) decreases with increasing distance from Syria and that, in each region, the %K tends to decrease with increasing time after the arrival of farming. Both the model and the data display a local minimum of the genetic cline, and for the same Neolithic regional culture (Sweden). Comparing the observed ancient cline of haplogroup K to the simulation results reveals that about 98% of farmers were not involved in interbreeding neither acculturation (cultural diffusion). Therefore, cultural diffusion involved only a tiny fraction (about 2%) of farmers and, in this sense, the most relevant process in the spread of the Neolithic in Europe was demic diffusion (i.e., the dispersal of farmers), as opposed to cultural diffusion (i.e., the incorporation of hunter-gatherers).

The Neolithic transition was a major transformation that introduced agricultural economics, radically changed the environment, and led to increased population densities and new forms of social organization[1,2]. The Neolithic spread from the Near East across Europe, from about 8,000 yr Before the Common Era (BCE) until about 3,000 yr BCE[3]. A crucial question is whether the spread of the Neolithic was due to a dispersal of farming populations (demic diffusion), to the learning of agricultural techniques by European hunter-gatherers (cultural diffusion), or to a combination of both mechanisms. The latter possibility is suggested by the comparison of archaeological data to mathematical wave-of-advance models, which indicate that demic diffusion was more important than cultural diffusion[4,5]. Genome-wide studies also indicate a crucial role for demic diffusion, with very little cultural diffusion at the beginning of the Neolithic[6,7]. Notwithstanding the unquestionable importance of genome-wide studies, it is also of interest to analyze specific genetic markers, for two reasons. First, genome-wide studies cannot provide any quantitative explanation for the observed spatial cline of a single marker. And secondly, genome-wide studies cannot yield a quantitative estimate for the percentage of farmers involved in cultural diffusion. In order to understand both limitations of genome-wide studies, consider first one marker that has not been affected by drift neither selection. If there is admixture between the populations of incoming farmers and indigenous hunter-gatherers (HGs), and the latter originally lacked this marker, then it will dilute progressively, i.e. its frequency will decrease with increasing distance from the spatial region of origin of the Neolithic front. Second, consider again a maker unaffected by drift neither selection, but such that HGs initially had higher frequencies than farmers. Its frequency will not decrease but increase with distance from the Neolithic origin. Thirdly, consider a marker that increased its frequency after some location during the spread of the Neolithic front (due, e.g., to surfing or other drift effects). If HGs originally lacked this marker, its frequency will decrease (due to admixture) up to some distance, and increase for larger distances. Fourthly, if several drift and/or selective effects took place, the cline can have even more complicated shapes. Thus, clearly the frequencies of different genetic markers have different spatiotemporal dependencies, because they are due to different processes. For this reason, in order to estimate the percentage of farmers involved in cultural diffusion we should not to include many arbitrary markers (as in genome-wide studies). Instead, we should consider very specific markers that satisfy the following conditions: (1) the frequency decreases with increasing distance from the spatial origin of the Neolithic front; (2) HGs

[1]Complex Systems Laboratory, Universitat de Girona, C/Maria Aurèlia Capmany 61, 17003, Girona, Catalonia, Spain. [2]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 3, 08010, Barcelona, Catalonia, Spain. Correspondence and requests for materials should be addressed to J.F. (email: joaquim.fort@udg.edu)

lack the marker considered before the arrival of the first farmers (otherwise we would need to know the precise space-time variation of the marker initial frequency in HGs); (3) selection and (4) drift (including surfing) effects can be neglected. This makes it possible to compare the data to demic-diffusion models neglecting drift, selection, etc. (as done below). In the present paper we analyze mitochondrial haplogroup K because, as we shall see, the observed data for this marker satisfy conditions (1) and (2). In contrast, other markers that have been found in Early Neolithic European sites (e.g., N1a, J, T and X) have not been found in the Near East[8], so condition (1) does not hold. Condition (3) can be also reasonably assumed, because there are no data indicating the existence of any selective pressure on haplogroup K, and analysis of the Early Neolithic K haplotypes does not show signs of selection (Supplementary Text S1). It is also reasonable to assume that condition (4) holds, because we will show that a simulated cline (neglecting drift) is consistent with the observed one for haplogroup K.

A totally independent reason why genome-wide studies cannot determine quantitatively the percentage of farmers involved in cultural diffusion is that, e.g., Mathieson *et al.*[6] assume only two source populations, namely an Anatolian Neolithic one and Western HGs, and use f_4-statistics to estimate, e.g., a 93% of Anatolian Neolithic ancestry and a 7% of Western HG ancestry for Early Neolithic farmers in Germany. But this result of 93% is not the *percentage of farmers involved in cultural diffusion*. Instead, it is the Anatolian fraction ($\alpha_1 = 0.93$) of genetic drift (defined as a variance of allele frequencies[9]) of the German population considered (assuming that its drift is a linear combination of the drifts of the two presumed source populations). But there is no mathematical theory relating the proportions of genetic drift (i.e., the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_N$ of the f_4-value of a test population in terms of the f_4-values of its N presumed source populations[6,9]) to the percentage of farmers involved in cultural diffusion. Similarly, in admixture analysis the fractions of the genome contributed by a set of presumed source populations are estimated, but again there is no theory relating them to the percentage of farmers involved in cultural diffusion. For totally analogous reasons, these and other previous methods (f_4-statistics, admixture, principal components, structure analysis, D-statistics, etc.) can provide valuable qualitative indications on whether demic or cultural diffusion dominated the Neolithic spread, but they cannot yield any quantitative value for the percentage of farmers involved in cultural diffusion. Incidentally, we note that many such methods (e.g., f_4-statistics and admixture) assume a few source populations, whereas here we will consider the more realistic case of populations distributed continuously in space (and also include the effect of seas and mountains). If clinal patterns are not observed in analyses based on principal components, admixture, f_3, f_4, D-statistics, etc. (where, instead, early Neolithic individuals tend to cluster together, e.g. with modern Sardinians), the reason is simply that those analyses are based on many markers but, as explained above, the spatial distribution of each one can be due to other processes in addition cultural diffusion (surfing, other kinds of drift, selection, etc.).

In this article we shall estimate the percentage of early farmers involved in cultural diffusion from an ancient DNA (aDNA) marker. We will perform our analysis at the continental scale, because aDNA data are not yet numerous enough to consider specific geographic regions. As we shall see, however, there are already sufficient data to obtain some first estimates of general trends. We consider mitochondrial DNA (mtDNA), because nuclear data are known for a substantially smaller number of ancient individuals. Mitochondrial DNA is inherited from the mother, thus its study will be related to the spread of maternal lineages. As all genetic sequences, mtDNA can be inherited with mutations, but similar sequences (haplotypes) with a common ancestor are usually grouped into haplogroups. Since the aDNA data are still limited in number, we perform our analysis below at the haplogroup level, grouping together the different haplotypes and subclades from each lineage (in Supplementary Text S1 we include analyses at the haplotype level, and they reinforce our conclusions). The mtDNA of European hunter-gatherers is composed mainly of U lineages (U, U4, U5, and U8), which are absent in early Neolithic populations[10–12]. Conversely, haplogroups N1a, T2, K, J, HV, V, W, and X have been proposed as potential Neolithic markers because they have been found in farmers of the Linearbandkeramic (LBK) culture, an early Neolithic culture in Central Europe, and are almost absent in neighboring hunter-gatherer samples[10,13]. Haplogroup K has been identified in only three hunter-gatherers dated before the arrival of farming (two in Greece[14] and one in Georgia[12]), but their subclades have not been found so far in any Neolithic farmer (see Supplementary Text S2 for a detailed discussion of the very few exceptions of Mesolithic individuals displaying K haplotypes). Thus haplogroup K was virtually absent in Europe before the arrival of farming, and condition (2) above is satisfied. On the other hand, as we shall see below, haplogroup K displays a cline of decreasing frequency with increasing distance from the spatial origin of the Neolithic expansion. Thus haplogroup K also satisfies condition (1) above, in contrast to other potential Neolithic markers (N1a, T2, J, HV, V, W, and X).

Results and Discussion

In order to study the existence of a genetic cline for haplogroup K in early Neolithic populations and subsequently compare it to our simulations, we have gathered all mtDNA information of Early and Middle Neolithic individuals reported in the literature, and we have grouped the data into regional cultures according to their location, date and reported culture (Supplementary Data S1). The Neolithic expansion in Europe begun in the Near East, and for this reason we have used the oldest pre-pottery Neolithic B (PPNB) date from Syria[3], Ras Shamra, as a geographic reference for the origin of the spread. In Fig. 1 we represent, for each regional culture, the average date of its individuals whose mtDNA haplogroup has been determined against the distance from their average location to Ras Shamra. Figure 1 includes all regional cultures dated between the Early and the Middle Neolithic, such that the mtDNA haplogroup of more than two individuals is known (e.g., Greece could not be included; see Supplementary Text S3). The Southern Levant is not included for reasons explained in Supplementary Text S3. For each regional culture, the number of individuals whose mtDNA haplogroup has been determined is given in the caption to Fig. 1 (Supplementary Data S2–S3). We distinguish 3 different groups of regional cultures in Fig. 1. The first group is composed by the 10 *oldest* Neolithic regional cultures (from Syria to western and northern Europe) for which there are genetic data (squares in Fig. 1). The second group (triangles in Fig. 1) corresponds to 15 regional cultures that have *younger* dates than the oldest ones (squares) and that are located at similar distances

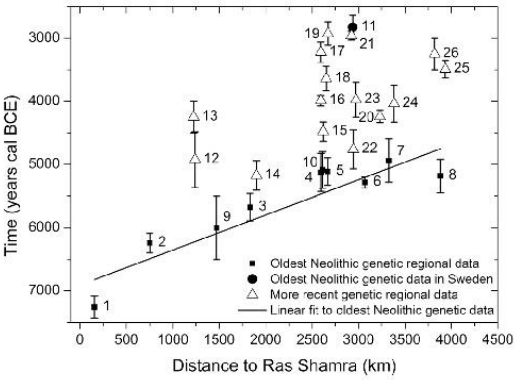


Figure 1. Dates versus great-circle distances from Ras Shamra (Syria) for 26 regional cultures with ancient mtDNA data. Squares correspond to the oldest regional Neolithic cultures, namely 1 Syria PPNB (15 individuals), 2 Anatolia (28 individuals), 3 Hungary-Croatia Starčevo (44 individuals), 4 Eastern Germany LBK (36 individuals), 5 Western Germany LBK (56 individuals), 6 Northeastern Spain Cardial (15 individuals), 7 Spain Navarre (36 individuals), 8 Portugal coastal Early Neolithic (10 individuals), 9 Romania Starčevo (5 individuals) and 10 Southern Germany LBK (4 individuals). The circle stands for 11 Sweden (9 individuals), which is substantially delayed due to the slowdown of the Neolithic front in northern Europe. Triangles correspond to more recent regional cultures, namely 12 Romania Middle Neolithic (29 individuals), 13 Romania Late-Middle Neolithic (9 individuals), 14 Hungary LBK (45 individuals), 15 Eastern Germany RSC (10 individuals), 16 Eastern Germany SCG/BAC (38 individuals), 17 Eastern Germany SMC (30 individuals), 18 Western Germany BAC (14 individuals), 19 Western Germany BEC (17 individuals), 20 Western France Prissé (3 individuals), 21 South-Eastern France Treilles (29 individuals), 22 Catalonia Epicardial (7 individuals), 23 Catalonia Late Epicardial (3 individuals), 24 Spain Basque country (7 individuals), 25 Portugal coastal Late Neolithic (3 individuals) and 26 Portugal inland Late Neolithic (7 individuals). The straight line is the regression fit to the 10 oldest regional data (squares). The symbols and extremes of each error bar give the averages of the mean, maximum and minimum calibrated dates, computed over all individuals with known mtDNA in the corresponding regional culture (Supplementary Data S1).

from Syria (i.e., broadly in the same area). Thus, the triangles in Fig. 1 are not representative of the earliest local Neolithic cultures. Finally, the circle in Fig. 1 corresponds to Sweden. Its date and location are those of the earliest Neolithic individuals in Sweden whose mtDNA is known. It would be thus legitimate to consider this data point (circle in Fig. 1) simply as one of the oldest regional cultures (squares), and we will actually include it into our calculations below. But the date for Sweden is substantially delayed relative to other cultures located at similar distances (Fig. 1), so it will be useful to identify Sweden with a symbol (circle) different than the other oldest regional cultures (squares).

Understanding the observed variations in the percentage of haplogroup K (%K). It is important to keep in mind that the *oldest* regional cultures displayed in Fig. 1 do not correspond to the oldest archaeological dates known for each Neolithic regional culture, but only to the oldest Neolithic individuals whose mtDNA haplogroup has been determined. In spite of this, those dates (squares in Fig. 1) show a highly linear dependence on distance (correlation coefficient $R = 0.93$), as predicted for the oldest dates by wave-of-advance models[4]. In Fig. 2 we plot the %K as function of distance from Ras Shamra (Syria) for all the regional cultures in Fig. 1 that include at least 8 individuals (regions with fewer individuals have been ignored to avoid very large error bars). Because the total number of individuals per region is still small in many regions, in our analysis below we take into account the whole 80% confidence-level (80% CL) range, represented as error bars, rather than only mean values. Labels and symbols in Fig. 2 are the same as in Fig. 1. For the oldest Neolithic cultures, there is no theoretical reason to expect a linear dependency of the %K on distance (in other words, we should not expect a high value of R for the regression to the squares and circle in Fig. 2). However, the slope of this regression in Fig. 2 is highly significantly different from zero ($P = 0.001$), and this gives statistical support to the existence of a genetic cline (similarly, low values of P also yield statistical support to the existence of phonemic clines[15,16]). Additional support to the existence of a cline is obtained from an interpolation map and the analysis of the %K data by means of a Moran's I correlogram, included in Supplementary Text S4.

As explained in the Introduction, it has been proposed that haplogroup K spread across Europe from the Near East with the Neolithic front[8,10,17–19]. We shall call this proposal the wave-of-advance model of haplogroup K. The analysis of the Early Neolithic haplotypes in our database also yields support to the assumption that the population with haplogroup K underwent a recent process of demographic and geographic expansion (Supplementary Text S1). Obviously demic diffusion, on its own, cannot explain the spatiotemporal distribution of haplogroup K (as displayed in Fig. 2), because purely demic diffusion predicts a uniform distribution (see (i) below and Sec. Demic versus cultural diffusion). Thus we ask whether cultural (in addition to demic) diffusion is a viable

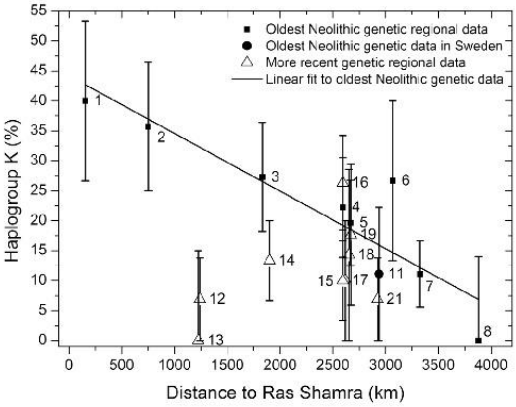


Figure 2. Observed percentage of mtDNA haplogroup K as a function of the great-circle distance from Ras Shamra (Syria). Each number denotes the same culture as in Fig. 1 (regions with fewer than 8 individuals have been ignored to avoid very large error bars). The straight line is the regression fit to the 10 oldest regional data (squares) and the oldest data in Sweden (circle). Error bars display 80% CL intervals (see Materials and Methods, Statistical analysis and Supplementary Text S10).

explanation. If HGs lacked haplogroup K (as justified by genetic data in the Introduction and Supplementary Text S2), and other effects (selection, drift, mutation, etc.) can be neglected, a demic-cultural model makes two testable predictions. (i) For the earliest Neolithic cultures, we should observe a decrease in the percentage of farmers with haplogroup K with increasing distance from the Near East (because interbreeding of pioneer farmers with local hunter-gatherers, and/or acculturation of the latter during the front propagation, will diminish the %K). This prediction is clearly observed in Fig. 2, because the earliest regional Neolithic cultures (squares and circle) show a clear decrease of the %K with increasing distance from Syria. (ii) For each region, this model also predicts that the earliest Neolithic regional culture will have a higher percentage of farmers with haplogroup K than later cultures (due to interbreeding and acculturation subsequent to the arrival of the Neolithic wave of advance). Prediction (ii) is also observed in Fig. 2, because of the 9 European cultures that do not correspond to the earliest Neolithic (triangles), only 1 (culture 16) has a larger %K than the expected regional maximum (the latter is given by the linear fit to the earliest regional Neolithic cultures in Fig. 2), and even culture 16 may be lower than the expected maximum, if the error bar is taken into account. However, we must caution that prediction (ii) refers to populations dated substantially later than the spread of the Neolithic front and it is therefore affected by population movements and other processes subsequent to the spread of the Neolithic. Thus, it is not reasonable to try to explain *quantitatively* prediction (ii) with a simulation model of the spread of the Neolithic. For this reason, although the model satisfies *qualitatively* both predictions, in the rest of this paper we shall be mainly concerned with prediction (i).

Ancient cline of haplogroup K. Figure 3 shows (lines) the clines obtained from our wave-of-advance simulations of the Neolithic and haplogroup K spread (see Materials and Methods and Supplementary Texts S5–S9), alongside the observed genetic data for the earliest regional Neolithic cultures (squares and circle) already depicted in Fig. 2. In Fig. 3 we have imposed the initial genetic conditions that all simulations predict the observed %K for Syria (square labelled 1; see more details on the implementation of the initial conditions in Materials and Methods and Supplementary Text S7). The simulated clines have been computed at the same 9 locations and dates as the genetic data (so the lines simply join the 9 data points), and for several values of the cultural diffusion intensity η.

We first observe that, similarly to the behavior of the data (symbols in Fig. 3) and in agreement with prediction (i) formulated above, when considering cultural diffusion ($\eta \neq 0$), the %K from the simulations (lines) tends to decrease with increasing distance from the Near East. This behavior was to be expected, because more distance from the origin (Ras Shamra, Syria) implies more time for the farming populations to interact (via interbreeding or/and acculturation) with hunter-gatherers (who lack haplogroup K). However, we note that both the simulations and the data display a local minimum at region 11 (Sweden). This is due to the fact that, according to archaeology[3] and ancient genetics[20,21], the spread of the Neolithic in Europe occurred following two main routes: one along the Mediterranean coast (corresponding to the Impressa and Cardial traditions) and the other through the Balkans and the Central European plains (corresponding to the Starčevo and LBK cultures). To see how this explains the minimum in Fig. 3, consider first the Neolithic front propagating along the Mediterranean coast. In this case, population dispersal is driven by jumps (maritime migrations) of about 150 km per generation (as shown in Materials and Methods and Supplementary Text S6, and in agreement with previous simulation results[3]). Conversely, the Neolithic front propagating inland is driven by jumps of about 50 km per generation (Materials and Methods). Therefore, in order for the Neolithic front to travel a given distance, a *coastal*

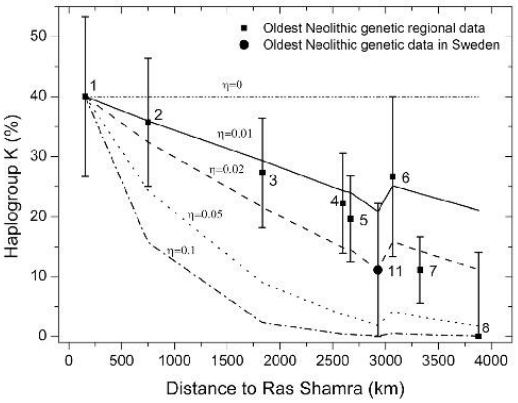


Figure 3. Observed and simulated percentage of mtDNA haplogroup K as a function of the great-circle distance from Syria. The data are shown with the same error bars as in Fig. 2, but only for the oldest regional cultures. The lines are the results of the mathematical simulation for several values of the cultural diffusion intensity η. The lines have been plotted by joining the simulation results for each of the 9 regional cultures, obtained at the average location and date of the individuals whose mtDNA haplogroup has been determined for each regional culture (Supplementary Data S1). Therefore, the simulation result for each region has been obtained at its average date (Fig. 1 and Supplementary Data S1). Numerical labels denote the same cultures as in Figs 1–2.

propagation obviously implies fewer jumps, i.e., fewer generations, and therefore less time for interbreeding with hunter-gatherers (and/or acculturation of the latter) than an *inland* propagation. Thus a *coastal* route will lead, at a given distance, to a lower decrease of the %K than an *inland* route. This is why the Mediterranean route leads, in region 6 (NE Spain) in Fig. 3, to higher values of the %K than the central-northern European route in region 11 (Sweden), in spite of the fact that the former is further away from Syria than the latter. This explains the minimum in the simulation curves (and in the observed data) in Fig. 3 (see Supplementary Text S8 for a more detailed discussion, and Fig. S15 for a plot of the simulated clines along both routes).

Demic versus cultural diffusion. What does the observed cline of haplogroup K for Early Neolithic cultures (error bars in Fig. 3) imply about the importance of cultural diffusion in the spread of the Neolithic? First, let us examine how the intensity η of cultural diffusion is related to the steepness of the genetic cline. Note that, in the absence of cultural diffusion (i.e., without interbreeding neither acculturation), the %K at all farming populations would remain approximately constant at the value observed for the original (PPNB) population in Syria (assuming that drift and other processes do not have a strong effect). Thus, in a purely demic model ($\eta=0$), such a cline would not be observed. Accordingly, the simulation for $\eta=0$ leads to a uniform distribution in Fig. 3. We also expect that the stronger the intensity of cultural diffusion, the more important the decrease in the frequency of haplogroup K, and the steeper its geographic cline. This intuitive expectation agrees with the simulation results in Fig. 3, where for any given distance from the origin, a higher value of the cultural transmission intensity η yields a lower %K.

By comparing the data (symbols) to the demic-cultural space-time simulations (lines), we observe that Fig. 3 implies that the intensity of cultural diffusion was $\eta \approx 0.02$ (because higher or lower values of η lead to lines that are not within all of the error bars obtained from the aDNA data). The maximum possible value of this parameter is $\eta = 1$[22] (see the text below equation (2)). Therefore, although the observed cline cannot be explained without cultural diffusion ($\eta=0$, horizontal line in Fig. 3), such a low value ($\eta \approx 0.02$) implies that cultural diffusion was remarkably weak. Indeed, the cultural diffusion intensity η can be interpreted as the proportion of pioneering farmers that mate a hunter-gatherer[22] or, alternatively, that teach agriculture to a hunter-gatherer[4] (Supplementary Text S9). Thus, our result that $\eta \approx 0.02$ (Fig. 3) implies that cultural diffusion involved only a tiny fraction (about 2%) of farmers and, in this sense, the most relevant process in the Neolithic spread in Europe was demic diffusion. Modifying the initial conditions so that the whole 80% CL for Syria is considered refines this estimate of the percentage of farmers involved in cultural transmission to the range $2\% \pm 1\%$ (Supplementary Text S7). The primacy of demic diffusion has been noted in genome-wide studies (see, e.g., previous work by Mathieson *et al.*[6]), but those studies could not quantify the percentage of farmers involved in cultural diffusion (see our Introduction). In contrast, we quantify that about 98% of farmers did not take part in cultural diffusion.

Our main result, namely that a very small amount of cultural transmission is enough to produce a continent-wide genetic cline, agrees with previous simulations[23–25], which however did not use the equations of cultural transmission theory recently derived[4,22] (see equations (1)-(3) and Supplementary Texts S5, S9 and S11) nor could compare to aDNA data (which were then also unavailable). Therefore, in none of those previous studies was it possible to estimate quantitatively the percentage of farmers involved in cultural diffusion.

Conclusions

In this paper we have analyzed the genetic implications of a mathematical model that combines demic dispersal, population growth, and cultural transmission theory. Using anthropologically realistic assumptions and parameter values, we have performed, to the best of our knowledge, the first qualitative and quantitative comparison of a mathematical model to an observed Neolithic genetic cline. Although the ancient genetic data currently available are still limited, especially those corresponding to the Early Neolithic, they cover a wide enough area (see Supplementary Text S4, Fig. S7) to allow us to analyze the geographical cline of genetic markers at the continental level, even if regional variations cannot be detected. In addition, the data are numerous enough so that we can observe a cline, and reach conclusions valid at least at the 80% CL (error bars in Fig. 3 and in Supplementary Text S7, Figs S12–S14). A Moran's I correlogram confirms the existence of the cline (Supplementary Text S4, Fig. S8). We have focused our attention on haplogroup K, mainly because it is virtually absent in hunter-gatherer populations and its frequency has a maximum in the Near East (specifically in Syria). Both points make it possible to attempt a description based on a simple mathematical model.

Qualitatively, the model predictions agree with the data in two ways: (i) both the data and the simulations show that the %K tends to decrease with increasing distance from Syria (Fig. 3); (ii) for each region, the %K tends to decrease with increasing time after the arrival of farming (Fig. 2).

Quantitatively, comparison between the model and the data shows that: (i) both the model and the data display a local minimum of the genetic cline, and for the same regional culture (Sweden, i.e. symbol 11 in Fig. 3); (ii) the ancient cline of haplogroup K can be explained if about 98% of farmers were not involved in cultural diffusion. However, we stress that the observed cline cannot be understood assuming that 100% of farmers were not involved in cultural diffusion. Thus, the observed cline implies that some farmers took part in cultural transmission (either by interbreeding or by teaching agriculture to hunter-gatherers). But only a tiny fraction (about 2%) of farmers were involved in cultural diffusion. In this sense, the most relevant process in the expansion of Neolithic culture in Europe was demic diffusion, i.e. the reproduction and dispersal of farmers, as opposed to the incorporation of hunter-gatherers (cultural diffusion).

Recently, the conclusion that the spread of the Neolithic in Europe was driven mainly by demic diffusion has been also obtained from comparing non-genetic, demic-cultural models to the spread rate of the Neolithic front, as estimated from archaeological data[4]. However, using only archaeological data has severe limitations. The reason is the following. Archaeological data make it possible to estimate the spread rate of the Neolithic wave of advance, and this can be compared to the results of the mathematical model. But the dependence of the spread rate on the intensity of cultural transmission is weak[4,22] and, for this reason, the spread rate can be used only to estimate an upper bound for the intensity of cultural transmission (namely $0 < C < 2.5$[4], equivalent to $0 < \eta < 2.5$ here, see Supplementary Text S9). In contrast, here we have shown that genetic data make it possible to know a function that depends strongly on the intensity η of cultural transmission (Fig. 3), namely the percentage of the considered haplogroup as a function of distance (i.e., the genetic cline shown in Fig. 3). This strong dependency has made possible a much more precise estimation of the percentage of farmers involved in cultural diffusion, namely $\eta = 0.02$ (Fig. 3), i.e. about 2%. This shows the tremendous potential of combining genetics, archaeology and mathematical modelling. On the other hand, the high number of archaeological data has allowed the identification of regional variations[5], something that is still not possible on the basis of ancient genetic data.

Our findings agree with genome-wide results, in the sense that demic diffusion was the main driver of the Neolithic spread in Europe (see, e.g. the results by Mathieson *et al.*[6]). However, genome-wide studies cannot estimate the percentage of farmers involved in cultural diffusion (see our Introduction). In contrast, our methodology yields the first quantitative estimation for this percentage (about 2%). This is possible because, in contrast to genome-wide studies, our approach has two crucial features: first, we compare to cultural-demic wave-of-advance mathematical models; second, we use a marker that shows decreasing frequency with increasing distance from the Near East. This estimate arises from comparing our model to the data at the 80% CL, leading to a confidence interval for the importance of cultural diffusion of $2\% \pm 1\%$. Of course, if additional such markers are identified in future work, they will yield more precise results and will also allow the study of regional variabilities. Thus the present paper is a first step, which also provides a plausible explanation for the observed cline of haplogroup K at a continental scale. We stress that such an explanation cannot be provided by genome-wide studies. For simplicity, our models assume the same dispersal behavior for males and females. If future studies detect ancient clines of decreasing frequency for additional genetic markers, and they consistently show differences between maternal and paternal markers, they could be used to infer different dispersal behaviors for females and males, using trivial extensions of our models.

Ancient DNA data indicate that cultural diffusion was more important in some specific regions, such as Scandinavia[26] or the Paris Basin[27]. Thus, it has been recently suggested that the effect of cultural diffusion increased as farmers migrated farther west in Europe[27]. This suggestion agrees nicely with: (i) our simulated clines (lines in Fig. 3); (ii) the observed cline of haplogroup K (symbols in Fig. 3); and (iii) the intuitive expectation that longer distances from the spatial origin of the Neolithic imply more time for interbreeding and/or acculturation and, therefore, a stronger effect of cultural diffusion.

Materials and Methods

Archaeological and genetic data. We gathered a database of all individuals from farming cultures dated between 8,000 and 3,000 calibrated years BCE for which the mtDNA haplogroup have been reported in the literature. For all 513 individuals in the database, we report the haplogroup, date, latitude, longitude, bibliographical references and additional data (Supplementary Data S1). We grouped them into regional cultures according to their geographical and cultural closeness (*e.g.*, Syria PPNB, Anatolia, Hungary-Croatia Starčevo, Hungary LBK, etc.). The data from Syria are from PPNB sites, which makes them especially relevant because PPNB/C

are the Near-Eastern Neolithic cultures that later spread into Europe[3]. We selected for further analysis the 26 regional cultures with more than two individuals (comprising 508 individuals), and discarded the others (see Supplementary Text S3 for a discussion on Neolithic individuals not included in the analysis). For each of the 26 selected regional cultures, we calculated the percentage of individuals with K haplotypes (Supplementary Data S2–S3), the average date of its individuals, and the average great-circle distance of its individuals to the site of Ras Shamra (Supplementary Data S3). This is the oldest PPNB Syrian site used in previous simulations studies[3], and we therefore use it as a reasonable geographic reference from the origin of the Neolithic range expansion in our simulations (see below).

Statistical analysis. For each of the 26 regional cultures, we estimated the error intervals of its average date and %K. The time error bar (Fig. 1) was estimated by averaging the reported maximum and minimum dates for all individuals in the considered regional culture whose mtDNA is known. The error bar for the %K (Figs 2–3) was estimated by the bootstrap method, computing the 80% CL interval of 10,000 replicates, except for the two regions where none of the sampled individuals have haplogroup K ('Portugal coastal Early Neolithic' and 'Romania Late-Middle Neolithic'). Then the bootstrap method cannot be applied directly (because the error would be exactly zero, which is not reasonable), and thus we applied a different statistical method, explained in detail in Supplementary Text S10. We have established the existence of the cline in 3 ways: linear regression (Fig. 2), interpolation map and Moran's I correlogram (Supplementary Text S4).

Analysis of K haplotypes. We have applied several statistical and phylogenetic analysis to the K haplotypes found in the 9 Early Neolithic regional cultures: we have computed Tajima's D and Fu's F_s neutrality tests; analyzed the geographical variation in the haplotype diversity, mismatch distributions, and first principal component; correlated genetic and geographical distances through Mantel test; performed network analysis; and constructed a Bayesian Skyline Plot (Supplementary Text S1). The obtained results show clear signs of a recent demographic and spatial expansion, in agreement with our assumption that haplogroup K spread with the Neolithic wave. These analyses have also shown as that, in principle, the regions displaying high values of %K are not the result of sampling individuals from a single family (see Supplementary Text S1, sec. 2) Haplotype diversity).

Space-time genetic simulations. We use a rectangular grid of square cells that covers the European continent, the Near East and part of Asia and Africa, with each cell classified as inland, coast, mountain or sea[3]. We use cells of 50 km × 50 km, since 50 km is the value corresponding to the mobility per generation according to ethnographic data of preindustrial populations[28]. At each cell we can have individuals of three populations: farmers who *have* haplogroup K, P_N; farmers who do *not have* haplogroup K, P_X; and hunter-gatherers, P_{HG} (no hunter-gatherer has haplogroup K). Each population would in principle include several different haplotypes, but since we are not interested in the evolution of any individual haplotypes, for simplicity the model used in the main paper does not consider any lower level subgroups. Below we describe the most important processes of the model, but we include a more detailed description in Supplementary Text S5.

Initial conditions. We applied the initial condition that at 8,233 yr BCE, the date of Ras Shamra (the oldest PPNB site in Syria from previous work[3]), all of the grid was empty of farmers except the cell that contains this site. In this cell, we set at 8,233 yr BCE the hunter-gatherer population density to zero, and the farmer population density to its saturation value ($P_{F\,\max} = 3,\ 200$ individuals/cell, from ethnographic data[3,25]). The PPNB Syrian archaeological and genetic data have different times and locations (the archaeological data is dated at 8,233 yr BCE and the genetic data at 7,258 yr BCE). For this reason, we have to set the %K at the cell containing Ras Shamra by trial and error so that the simulation yields the adequate value of the %K at the time and location of the genetic data in Syria (see details in Supplementary Text S7). In all grid cells (except for the initial one), the hunter-gatherer population is initially set at its saturation value ($P_{HG\,\max} = 160$ individuals/cell, from ethnographic data[25]), assuming that none of them has haplogroup K (see the Introduction and Supplementary Text S2).

Defining a generation as the mean age of the parents at the time one of their offspring is born (not necessarily the first), in simulations we use the mean value $T = 32$ yr obtained from ethnographic data[29]. Let t stand for the number of generations elapsed since the beginning of the simulation (8,233 yr BCE). For $t = 1,\ 2,\ 3\ldots$ we apply the following cycle of 3 steps (changing their order would yield the same results):

(1) ***Dispersal***. At each cell, we update the values of P_N and P_X by computing how many farmers of both kinds arrive at the cell from other cells. We do this, as in previous work[3,22,28], with a simple model in which, for each cell, a fraction p_e (which is called the persistence in demography) of the population of farmers (independently of their genes) stays at the cell, and a fraction $(1 - p_e)$ relocates to the four nearest neighbor cells, each receiving a fraction $(1 - p_e)/4$. We use the mean value $p_e = 0.38$ obtained from ethnographic data[28]. We expect that including a set of distances and probabilities would lead to similar results[4]. If one or more of the nearest neighbors are mountain cells, they cannot receive population and each of the remaining neighbors receives a higher fraction. If one or more neighbors are sea cells, the corresponding fraction of the population (that would move there) travels by sea, and is equally distributed among coast cells that can be reached by sea in straight lines of up to 150 km (this is the adequate distance to obtain agreement with archaeological data, as seen in Supplementary Text S6 and in previous work[3]). We do not update the number of HGs in each cell due to their dispersal, because exchange of HGs between saturated cells has no effect (since all HGs lack haplogroup K) and we assume that they do not disperse appreciably into cells in which their number has been lowered due to cultural transmission (see step 2 below).
(2) ***Cultural transmission***. This is the only step that was not included in our previous non-genetic simulations

on a real map of Europe[3], because they considered only purely demic models. There are 3 modes of cultural transmission[30]. Vertical transmission is due to interbreeding (i.e., cross-matings between farmers and HGs). Horizontal (oblique) transmission is due to learning of agriculture by HGs from farmers of the same (the previous) generation. The latter two modes can be combined in a single mathematical model, namely horizontal/oblique transmission[4]. Here we shall consider only vertical transmission for simplicity, but we would reach the same conclusions if we considered, instead, any combination of vertical and horizontal/oblique transmission (Supplementary Text S9).

After dispersal, in each cell there is a population of P_{HG} hunter-gatherers and $P_N + P_X$ farmers. To determine the population numbers of the new generation, we have to compute the matings that take place between and within those 3 population groups, and then apply the reproduction step. We assume that children of cross matings between farmers and HGs are farmers, in agreement with ethnographic observations[31,32]. The number of cross matings between HGs and each group of farmers is[22]

$$couples\ HN = \eta \frac{P_{HG} \cdot P_N}{P_{HG} + P_N + P_X}, \tag{1}$$

$$couples\ HX = \eta \frac{P_{HG} \cdot P_X}{P_{HG} + P_N + P_X}, \tag{2}$$

where $P_{HG} + P_N + P_X$ is the total population present at the cell, and parameter η is the intensity of interbreeding[22]. The case $\eta = 1$ corresponds to random mating. The case $\eta > 1$ corresponds to more cross matings than under random mating[22], which is not realistic for farmers and HGs according to ethnographic data[32,33] (moreover, $\eta > 1$ can lead to negative population numbers[22]). Therefore, in practice $0 \leq \eta \leq 1$. From equations (1)-(2) it is very easy to find the number of individuals P'_{HG}, P'_N, and P'_X who do not take part in HN neither NX matings. We can use them to compute the number of matings between *farmer* individuals of different genetic groups (i.e., between populations P'_N and P'_X) by using again vertical cultural transmission theory, and taking into account we have no reason to assume that farmers of a genetic group (i.e., with or without haplogroup K) will have a preference for (neither against) mating with farmers of the same genetic group. Thus we apply random mating ($\eta = 1$)[22] for matings between farmers,

$$couples\ NX = \frac{P'_N \cdot P'_X}{P'_N + P'_X}, \tag{3}$$

(3) ***Reproduction***. We apply the following rules. (i) Each couple will have $2R_{0,i}$ children, because $R_{0,i}$ (the net fecundity) is computed per parent and there are two parents per mating ($i = F, HG$). Ethnographic data indicate that the children of cross matings with one HG parent are farmers[31,32], thus we use $R_{0,HG}$ for HH matings and $R_{0,F}$ for HN, HX, NN, XX and NX matings. If the number of individuals computed for some population group, cell, and time step is larger than its corresponding maximum ($P_{F\ max}$ or $P_{HG\ max}$), then we set it to the corresponding maximum value (Supplementary Text S5, sec. 3. Reproduction). We expect that a logistic model would yield similar results. In our simulations we use, from ethnographic data, $R_{0,F} = 2.45$[34], indicating that after a generation, the size of the new population is 2.45 times the size of the parent population. We assume that $R_{0,HG} = 1$, i.e. that the HG populations have reached a stationary state and they do not grow in number (not even after some HGs mate into the farming community, because converted HGs will still need part of the cell space after they become farmers); we do not expect our conclusions to change for other reasonable values of $R_{0,HG}$. (ii) For each kind of mixed genetic mating (HN and NX), in our simplest model we assume that the mother belongs to P_N in 50% of the matings, whose children will also carry haplogroup K since mtDNA is inherited from the mother (i.e., a 50% of the total offspring of mixed genetic matings will belong to P_N). A more complicated model, assuming that mothers in HN and HX matings are always HGs (which is closer to ethnographic observations[32]) yields very similar results (Supplementary Text S11).

All the steps in the model are computed using real values for the population numbers. If we used a stochastic procedure to approximate them to integer values (at every cell, iteration, and process step), we expect that in average we would obtain the same results. We run our simulation program for 200 iterations (generations of 32 yr) for each set of parameter values, so that it covers the time from the start of the spread (Syria, 8,233 cal yr BCE) until the latest genetic data in the database (Sweden, 2,825 cal yr BCE; Supplementary Data S3). At each iteration we compute the number of HG, N and X individuals at each cell and record the latter two, so that we can compute the simulated %K (namely, $\frac{P_N}{(P_N + P_X)} \cdot 100$) and compare it to the observed one from the reported mtDNA data at each regional culture and its average date (Supplementary Data S3). This is done in Fig. 3.

References

1. Pinhasi, R., Fort, J. & Ammerman, A. J. Tracing the origin and spread of agriculture in Europe. *PLoS Biol.* **3**, e410 (2005).
2. Bocquet-Appel, J. P. & Bar-Yosef, O. *The Neolithic demographic transition and its consequences* (Spinger Berlin, 2008).
3. Fort, J., Pujol, T. & vander Linden, M. Modelling the Neolithic transition in the Near East and Europe. *Am. Antiq.* **77**, 203–220 (2012).

4. Fort, J. Synthesis between demic and cultural diffusion in the Neolithic transition in Europe. *Proc. Natl. Acad. Sci.* **109**, 18669–18673 (2012).
5. Fort, J. Demic and cultural diffusion propagated the Neolithic transition across different regions of Europe. *J. Roy. Soc. Interface* **12**, 20150166 (2015).
6. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
7. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
8. Fernández, E. *et al.* Ancient DNA analysis of 8,000 B.C. Near Eastern farmers supports an early Neolithic pioneer maritime colonization of mainland Europe through Cyprus and the Aegean Islands. *PLoS Genet.* **10**, e1004401 (2014).
9. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
10. Brandt, G. *et al.* Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science* **342**, 257–261 (2013).
11. Bramanti, B. *et al.* Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326**, 137–140 (2009).
12. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).
13. Haak, W. *et al.* Ancient DNA from European early Neolithic farmers reveals their Near Eastern affinities. *PLoS Biol.* **8**, e1000536 (2010).
14. Hofmanová, Z. *et al.* Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci.* **113**, 6886–6891 (2016).
15. Atkinson, Q. D. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* **332**, 346–349 (2011).
16. Fort, J. & Pérez-Losada, J. Can a linguistic serial founder effect originating in Africa explain the worldwide a phonemic cline? *J. R. Soc. Interface* **13**, 20160185 (2016).
17. Gamba, C. *et al.* Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. *Mol. Ecol.* **21**, 45–56 (2012).
18. Olalde, I. *et al.* A common genetic origin for early farmers from Mediterranean Cardial and Central European LBK cultures. *Mol. Biol. Evol.* **32**, 3132–3142 (2015).
19. Szécsény-Nagy, A. *et al.* Tracing the genetic origin of Europe's first farmers reveals insights into their social organizations. *Proc. R. Soc. B* **282**, 20150339 (2015).
20. Hervella, M. *et al.* Ancient DNA from hunter-gatherer and farmers groups from Northern Spain supports a random dispersion model for the Neolithic expansion into Europe. *PLoS One* **7**, e34417 (2012).
21. Sampietro, M. L. *et al.* Palaeogenetic evidence supports a dual model of Neolithic spreading into Europe. *Proc. R. Soc. B* **274**, 2161–2167 (2007).
22. Fort, J. Vertical cultural transmission effects on demic front propagation: Theory and application to the Neolithic transition in Europe. *Phys. Rev. E* **83**, 056124 (2011).
23. Rendine, S., Piazza, A. & Cavalli-Sforza, L. L. Simulation and separation by principal components of multiple demic expansions in Europe. *Am. Nat.* **128**, 681–706 (1986).
24. Barbujani, G., Sokal, R. R. & Oden, N. L. Indo-European origins: a computer-simulation test of five hypotheses. *Am. J. Phys. Anthropol.* **96**, 109–132 (1995).
25. Currat, M. & Excoffier, L. The effect of the Neolithic expansion on European molecular diversity. *Proc. R. Soc. B* **272**, 679–688 (2005).
26. Skoglund, P. *et al.* Genomic diversity and admixture differs for stone-age Scandinavian foragers and farmers. *Science* **344**, 747–750 (2014).
27. Rivollat, M. *et al.* Investigating mitochondrial DNA relationships in Neolithic Western Europe through serial coalescent simulations. *Eur. J. Hum. Gen.* **25**, 388–392 (2017).
28. Fort, J., Pérez-Losada, J. & Isern, N. Fronts from integrodifference equations and persistence effects on the Neolithic transition. *Phys. Rev. E* **76**, 031913 (2007).
29. Fort, J., Jana, D. & Humet, J. M. Multidelayed random walks: Theory and application to the Neolithic transition in Europe. *Phys. Rev. E* **70**, 031913 (2004).
30. Cavalli-Sforza, L. L. & Feldman, M. W. *Cultural transmission and evolution: A quantitative approach* (Princeton University Press, 1981).
31. Ammerman, A. J. & Cavalli-Sforza, L. L. *The Neolithic transition and the genetics of populations of Europe* (Princeton University Press, 1984).
32. Cronk, L. From hunters to herders: Subsistence change as a reproductive strategy among the Mukogodo. *Curr. Anthropol.* **30**, 224–234 (1989).
33. Early, J. D. & Headland, T. N. *Population dynamics of a Philippine rain forest people: The San Ildefonso Agta* (University of Florida Press, 1998).
34. Isern, N., Fort, J. & Pérez-Losada, J. Realistic dispersion kernels applied to cohabitation reaction-dispersion equations. *J. Stat. Mech. Theor. Exp.* **2008**, P10012 (2008).

Acknowledgements

This work has been partially funded by Ministerio de Economía, Industria y Competitividad (Grant FIS-2016-80200-P), Fundación Banco Bilbao Vizcaya Argentaria (Grant NeoDigit-PIN2015E), and an Academia award from the Catalan Institution for Research and Advanced Studies (to J.F.).

Author Contributions

J.F. conceived the research and devised the statistical method in Supplementary Text 10. N.I. and V.L.R. wrote the simulation codes. V.L.R. compiled the genetic database and prepared figures. N.I. performed the genetic analyses in Supplementary Texts S1 and S4. J.F., N.I. and V.L.R. wrote the paper and the Supp. Info.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-11629-8

Competing Interests: The authors declare that they have no competing interests.