TESI DOCTORAL UPF 2014

# Context dependent selection in molecular evolution.

**Inna S. Povolotskaya**

---

Thesis Advisor

**Dr. Fyodor A. Kondrashov**

Bioinformatics and Genomics Programme

Centre for Genomic Regulation (CRG)



Barcelona, December 2014

*To my parents, who supported me along the way.*

# Aknowledgments

This thesis would have never seen the light of day without help from many people. First and the most important, my family, who supported me from the very beginning. My mother by far deserves the most credit in me choosing this path. Without her dedication to giving me the best education at the early years and support later in life I would have probably quit and engaged in a "real job". My father was the first to show me how beautiful nature is and has always encouraged my passion and pursuit. My husband, Alex, has given me a great deal of support, by unconditionally believing in me. And finally my son, Andrey, has made me to learn to be much more productive to balance between work and family.

Fedya, my thesis advisor and my first mentor has greatly contributed to my emergence as a scientist. His support and trust allowed me to wonder around in the search for answers and questions, creating my own research agenda. Toby, during his time in the lab, was very much of a role model for me, encouraging me to continue even when the things were not going so well and very much of an "older brother", teaching me different tricks of being a successful scientist. And finally, people in the lab, Margo, Peter, Dinara, Carla have created a wonderful and stimulating working atmosphere and have shown to be invaluable on many occasions.

# Abstract

Epistasis, or genetic interactions between different mutations, is theoretically predicted to play a substantial role in such evolutionary processes as emergence of sexual reproduction and recombination, speciation, adaptive evolution. However, there is little experimental or statistical evidence of the ubiquity of epistatic interactions in nature. Here, we study long-term protein evolution and show that the constant independent selection model cannot describe rates and patterns of protein divergence: protein sequences diverge beyond theoretical limits and the rate of divergence is much slower than predicted. We show that protein evolution is best explained under the assumption of rapid turnover of fitness values associated with individual amino acids. We further extend this computational study and build a theoretical model to capture the effect of non-constant selection on molecular evolution.

# Resumen

Se ha predicho teóreticamente que la epistasis, es decir, las interacciones genéticas entre diferentes mutaciones, cumple un rol sustancial en procesos evolutivos, tales como la emergencia de la reproducción sexual, la recombinación, la especiación y la evolución adaptativa. Sin embargo, existe poca evidencia experimental o estadística de la ubicuidad de las interacciones epistáticas en la naturaleza. Aquí, estudiamos la evolución de las proteínas a largo plazo, y demostramos que el modelo constante de selección independiente, no es capaz de describir las tasas y patrones de divergencia encontrados en las proteínas: las proteínas divergen mas allá de los límites teóricos y la tasa de divergencia es mucho mas lenta que la esperada. A su vez, demostramos que la evolución de las proteínas se explica mejor bajo la suposición de un intercambio rápido entre los valores de eficacia biológica asociados con aminoácidos individuales. Mas aún, extendemos nuestro estudio computacional y construimos un modelo teórico que captura el efecto de la selección inconstante sobre la evolución molecular.

# Contents

# 1. Introduction

## 1.1 Genotype-to-Phenotype

One of the most fascinating unresolved questions in modern biology is prediction of the phenotype of an organism from its genotype. This extremely complicated task is relevant to many different research topics, ranging from fundamental research to applied medicine. In evolutionary biology, natural selection acts at the phenotypic level, while units of evolutionary changes, mutations, happen at the genotypic level and knowledge of the fitness landscape, the complete collection of all possible genotype-to-phenotype connections [Wright, 1932], is essential for determining dynamics and the forces underlying genotypic changes. In human genetics, prediction of the disease risk in complex hereditary disorders and identification of causal pathogenic mutations, is the most important unresolved question. In epidemiology, correct prediction of the flu strains with the highest fitness, the ones which will prevail the next season, and production of effective vaccines against those strain can save thousands of lives.

What makes this question so hard to answer? In the simplest additive model, if we knew fitness effects of every allele present in a particular genotype, we could simply take their superposition and thus estimate the fitness of the organism. All of the currently existing methods indeed use this approach, as in evolutionary biology [Davydov et al., 2010, Siepel et al., 2005], human genetics [Sunyaev et al., 2001] and epidemiology [Luksza and Lässig, 2014]. These methods reach certain level of accuracy of at maximum 70-90%, but other factors need to be considered in order to improve precision. It is now clear that the mapping between genotype and phenotype is complex, and most phenotypes are affected by environmental factors and intricate genetic interactions between alleles.

## 1.2 Epistasis

Genetic, or epistatic, interactions, are deviations from additive genetic effects of alleles on the phenotype. Outcome of a mutation is conditional on the genetic background where it appears and in combination the effect of several mutations is different from their separate effects. Sometimes, the negative effect of one mutation can be masked by the presence of another mutation, or conversely, two benign mutations in combination can have a severe effect [Cordell, 2002, Phillips, 2008]. Such cases are called sign epistasis, when the sign of the difference in fitness between mutant and wild type depends on other mutations.

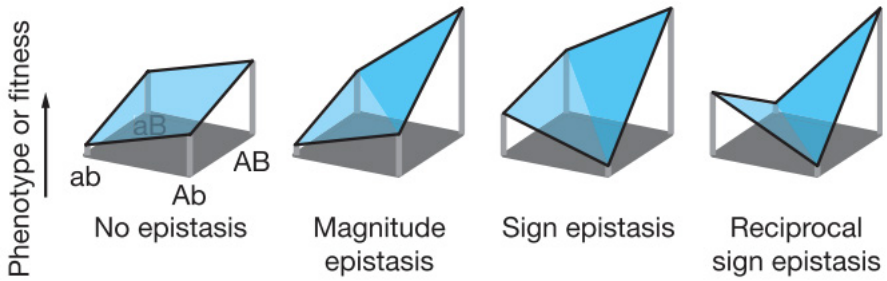


Figure 1.1: Different types of epistasis [Poelwijk et al., 2007].

In contrary, magnitude epistasis is a situation where mutations do not change the sign of the difference in fitness, but rather the absolute value of this difference. Two mutations are to interact antagonistically, if the difference in fitness of a double mutant is higher than the combination of their separate fitness effects and synergistically if the difference is lower.

Epistatic interactions play essential role in different evolutionary processes, as they determine the efficiency of natural selection in eliminating deleterious mutations from populations. Evolution of sexual reproduction, ploidy and recombination [Kondrashov and Crow, 1991, Kondrashov, 1988], speciation [Coyne, 1992, Dobzhansky, 1937, Kondrashov and Kondrashov, 1999, Orr, 1995], molecular evolution [Butcher, 1995, Kimura, 1985], protein divergence [Breen et al., 2012, Kondrashov et al., 2010, Povolotskaya and

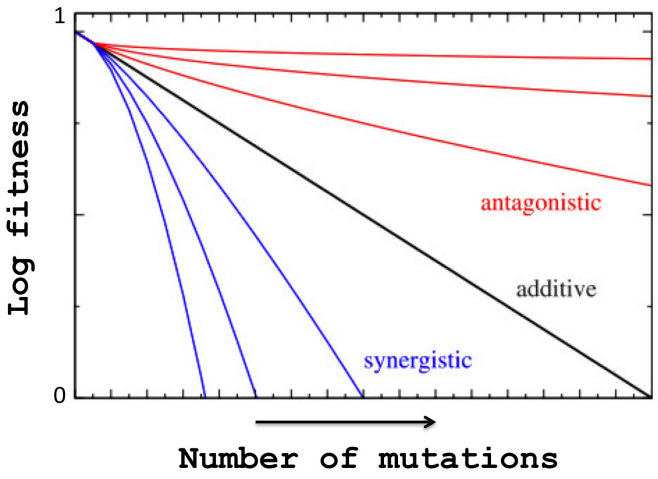


Figure 1.2: Red lines represent examples of antagonistic epistasis, the black line represents the null model of no epistasis and the blue curves represent cases of synergistic epistasis. Adapted from [Elena et al., 2010].

Kondrashov, 2010], adaptive evolution [Weinreich et al., 2006] all implicate presence of pervasive epistasis.

Here, I will give a brief overview of different studies of epistatic interactions performed to date. I will start with experimental conformations of presence of epistasis in different systems. Then, I will continue with the role of epistatic interactions in human genetics and identification of disease causing mutations. And at the end of current chapter I will overview the consequences of epistasis to protein evolution.

### 1.2.1 Experimental evidence of epistasis

Until recently, only sporadic cases of genetic interactions in natural populations were observed and described. One of the most prominent features of such data is its limitation to biased set of genotypes corresponding to high fitness. While these studies uncover presence of genetic interactions in a particular system, they can not be used neither for accurate reconstruction of genotype-to-phenotype maps, fitness landscapes, nor for estimating thorough distributions of the epistatic effects.

Naturally occurring genetic interactions are studied mostly for viral and bacterial species, due to the large number of available complete genomic sequences and relatively easy methods to measure fitness effects. One of the pioneering studies of epistatic interactions in human immunodeficiency virus 1, find statistical evidence for predominance of positive epistasis among substitutions [Bonhoeffer et al., 2004]. Studies in vesicular stomatitis virus [Sanjuán et al., 2004], in the RNA bacteriophage φ6 [Burch and Chao, 2004] and in the human H3N2 influenza virus [Gong et al., 2013], provided further evidence that antagonistic epistasis is an important characteristic of viral evolution.

In contrary, there is no clear pattern to the form of epistasis which prevails in bacteria. Studies, which focus on deleterious mutations, show that both synergetic and antagonistic epistasis are fairly common [Elena and Lenski, 1997, Maisnier-Patin et al., 2005]. Studies of the epistatic interactions between beneficial mutations reveal their antagonistic nature, consistent with the observation that the rate of adaptation to the new environment declines with time [Chou et al., 2011, Khan et al., 2011, Tenaillon et al., 2012, Wang et al., 2013]. One of the most striking examples is antibiotic resistance: alleles that confer antibiotic resistance interact antagonistically and the cost of multiple resistance is much smaller than expected under additive model [Trindade et al., 2009].

In eukaryotes, studies have shown rapid recovery of fitness from severe deleterious mutations by compensatory changes [Estes et al., 2011, Sza-

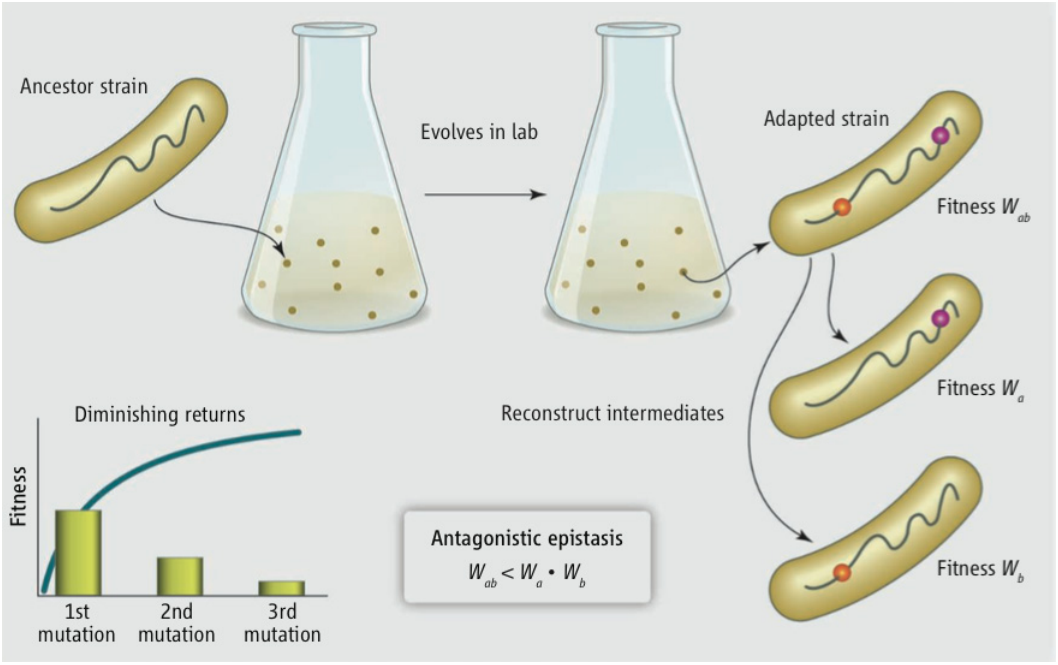


Figure 1.3: Antagonistic epistasis causes progressively slower rates of adaptation over time: mutations confer smaller marginal benefits in combination than they do individually. [Kryazhimskiy et al., 2011].

mecz et al., 2014]. Long-term mutation-accumulation lines of the nematode *Caenorhabditis elegans*, were maintained in large population sizes under competitive conditions. Fitness assays of these lines and comparison to parallel mutation-accumulation lines and the ancestral control showed that full recovery of mean fitness was achieved in fewer than 80 generations [Estes et al., 2011]. In a different study of 180 haploid baker's yeast genotypes displaying slow growth due to the deletion of a single gene authors have found compensatory evolution following gene loss to be rapid and pervasive: 68% of the genotypes reached near wild-type fitness through accumulation of adaptive mutations elsewhere in the genome. Genotypes with especially low fitnesses were more likely to be subjects of compensatory evolution. Genomic analysis revealed that as compensatory mutations were generally specific to the functional defect incurred, convergent evolution at the molecular level was extremely rare. Accordingly, compensatory evolution promoted genomic divergence of parallel evolving populations [Szamecz et al., 2014].

Overall picture of the patterns of epistasis across different organisms,

indicates correlation of average epistasis and genome complexity. In simpler genomes, such as those of viruses and some bacteria, interactions tend to be antagonistic. In unicellular eukaryotes, there seems to be no average deviation from independent effects, whereas in higher eukaryotes, a transition toward synergism occurs [Sanjuán and Elena, 2006].

However, in order to uncover a complete picture of fitness landscape, one would have to generate all possible mutants and look at their fitness effects distributions. It is extremely difficult task for separate mutations and multidimensional interactions, but it was performed for a number of organisms for pairwise interactions between genes. At the gene level, epistasis means that the phenotypic outcome of one gene depends on the level of expression of another one. Thus, when controlling expression of a gene, in both presence and absence of each of the other genes, it is possible to measure the amount of epistasis between them. In such a way, genetic interactions have been systematically mapped in *S. cerevisiae* [Costanzo et al., 2010], *C. elegans*[Byrne et al., 2007, Lehner et al., 2006], *D. melanogaster* [Horn et al., 2011] and in human cells [Bassik et al., 2013]. These studies has revealed enormous numbers: in the dataset which covered genetic interaction profiles for 75% of all genes in the budding yeast, [Costanzo et al., 2010] significant interactions were detected between ~170,000 different pairs of genes out of ~5.4 million pairs tested. One of the interesting features these studies have revealed a class of "hub" genes, which interact with other genes intensively.

While epistasis mapping by this approach has been extremely useful for detecting physiological connections between gene products, it is not well suited to investigate epistasis between mutations within a single gene. In order to uncover pairwise interactions in a particular protein, we need to screen for interactions between all possible amino acid combinations. For a protein of a length 100 amino acid, the number of such combinations would be $\frac{20^2 N!}{2!(N-2)!} \approx 10^6$. For now, there is no such method which could allow to routinely create a million of point mutants and contemporary studies are limited to small portion of such combinations.

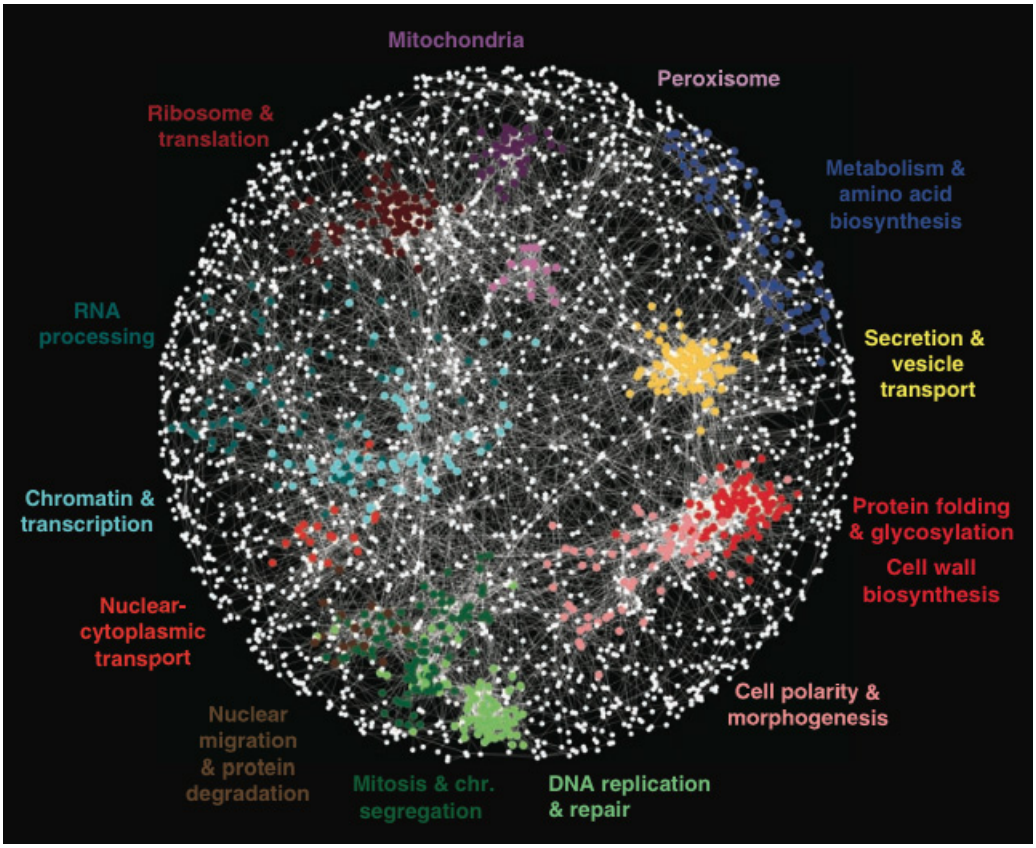


Figure 1.4: The complex network of genetic interaction in yeast [Costanzo et al., 2010].

One of the approaches is to look at the fitness effects of certain mutations, which correspond to the amino acid states in a different lineage. Such an experiment was performed by [Lunzer et al., 2010], The authors created 168 single mutations in wild-type *Escherichia coli* isopropymalate dehydrogenase (IMDH) that match the differences found in wild-type *Pseudomonas aeruginosa* IMDH. Surprisingly, only 104 mutant enzymes performed similarly to *E. coli* wild-type IMDH, one was functionally enhanced, while 63 were functionally compromised. Such a transition from *E. coli* IMDH, or an ancestral form, to the functional wild-type *P. aeruginosa* IMDH requires extensive epistasis to mask the combined effects of the deleterious mutations.

Another approach was developed recently, a method which can accurately measure the relative abundance of hundreds of mutations in a single bulk competition experiment, which can give a direct readout of the fitness of

each mutant [Hietpas et al., 2012]. Using random mutagenesis, an assessment of 40,000 double mutants in an RNA recognition motif in *S. cerevisiae* yielded the largest and most systematic picture of intragenic epistasis to date [Melamed et al., 2013]. Authors have found that while overall degree of pairwise epistasis is low, around 2%, depends on the proximity of the mutations in both primary and tertiary protein structure, and rises up to 10-15% for closely situated sites.

In another study [Bank et al., 2014], authors performed a fitness assessment of intragenic epistatic effects within a nine amino acid region of yeast Hsp90 implicated in substrate binding in seven Hsp90 point mutant backgrounds of neutral to slightly deleterious effect, resulting in an analysis of more than 1000 double-mutants. In contrary to the study [Melamed et al., 2013], authors found that negative epistasis between substitutions to be common (~46% of the tested pairs shoed epistasis), and positive epistasis to be rare, resulting in a pattern that indicates a drastic change in the distribution of fitness effects one step away from the wild type.

### 1.2.2 Epistasis in human genetics

One of the major problems, human geneticist are now facing, is the problem of so called "missing heritability". Genome-wide association studies have identified over 1200 genetic variants associated with over 165 complex human diseases and traits. Most variants identified so far confer relatively small increments in risk, and in combination explain only a small fraction of heritability. The "missing heritability", has been suggested to be explained by complicating factors such as an increased number of contributing loci and susceptibility alleles, incomplete penetrance, and contributing environmental effects. The presence of epistasis is a particular cause for concern, since, if the effect of one locus is altered or masked by effects at another locus, power to detect the first locus is likely to be reduced and elucidation of the joint effects at the two loci will be hindered by their interaction[Cordell, 2002, Manolio et al., 2009, Zuk et al., 2012].

In recent years there have been made a number of attempts to identify interacting single nucleotide polymorphisms (SNPs) which segregate in human population and contribute to complex diseases, but this task is challenging both experimentally and statistically as it would require sample sizes of an order of $10^6$ individuals.

On the other hand there is plentiful computational evidence that epistatic interactions play a key role in the recent evolution of the mammalian clade [Breen et al., 2012, Meer et al., 2010] and, in particular, human lineage [Baresić et al., 2010, Kondrashov et al., 2002, Soylemez and Kondrashov, 2012].

In a study [Kondrashov et al., 2002] authors investigated fitness landscape in the space of protein sequences by relating sets of human pathogenic missense mutations in 32 proteins to amino acid substitutions that occurred in the course of evolution of these proteins. On average, ~10% of deviations of a nonhuman protein from its human ortholog were identified as compensated pathogenic deviations (CPDs), i.e., were caused by an amino acid substitution that, at this site, would be pathogenic to humans. Normal functioning of a CPD-containing protein must be caused by other, compensatory deviations of the nonhuman species from humans. Later, [Soylemez and Kondrashov, 2012], broadened the analysis to 221 human genes with known pathogenic mutations to estimate the rate of irreversibility in protein evolution. Authors confirm previous result of approximately ~10% of all amino acid substitutions along the mammalian phylogeny being irreversible, and for a subset of 51 genes with high rates of irreversibility, as much as 40% of all amino acid evolution was estimated to be irreversible. As the fraction of irreversible states between two genomes is expected to increase with divergence as long as they represent cases of CPDs, authors estimated the fraction of irreversible states relative to genetic distance between humans and the common ancestor of six phylogenetically distant clades of placental mammals. Authors found a lower fraction of irreversible states in the common ancestors within the great apes consistent with the expectation under epistatic mode of evolution.

### 1.2.3 Evolution under epistasis

The concept of proteins evolving in a rigid protein space through the sequence of functioning intermediates was introduced by John Maynard Smith [Maynard Smith, 1970].

WORD WORE GORE GONE GENE

Figure 1.5: Word game as an analog of evolution through the functional intermediates [Maynard Smith, 1970].

In such cases, if a certain substitution is acceptable for evolution at the given moment, it will stay acceptable as long as some particular genetic background does not change. The probability that this genetic background has not changed is declining with time, meaning that under an epistatic model we expect the natural selection acting at the site to be time dependent, opposite to the conventional null hypothesis, when the selection remains constant over time [McCandlish et al., 2013]. Indeed, initially high rate of convergent evolution [Bazykin et al., 2007, Rokas and Carroll, 2008] declines with evolutionary distance between two organisms [Povolotskaya and Kondrashov, 2010, Rogozin et al., 2008] along with fitness of a newly emerged amino acid state [Naumenko et al., 2012].

If fitness effects of mutations are additive, they can happen in any order resulting in a monotonically increasing trajectory in fitness landscape towards a global maximum. Conversely, if mutations interact with one another by sign epistasis, the fitness landscape may become rugged and the fittest genotypes may only be accessed by accumulating mutations in one specific order. This makes it more likely that organisms will get stuck at local maxima in the fitness landscape [Poelwijk et al., 2011].

Rugged, epistatic fitness landscapes also affect the directionality of evolution. When a mutation has a large number of epistatic effects, each accumulated mutation drastically changes the set of available beneficial mutations and therefore an evolutionary trajectory, taken by a particular sequence, is

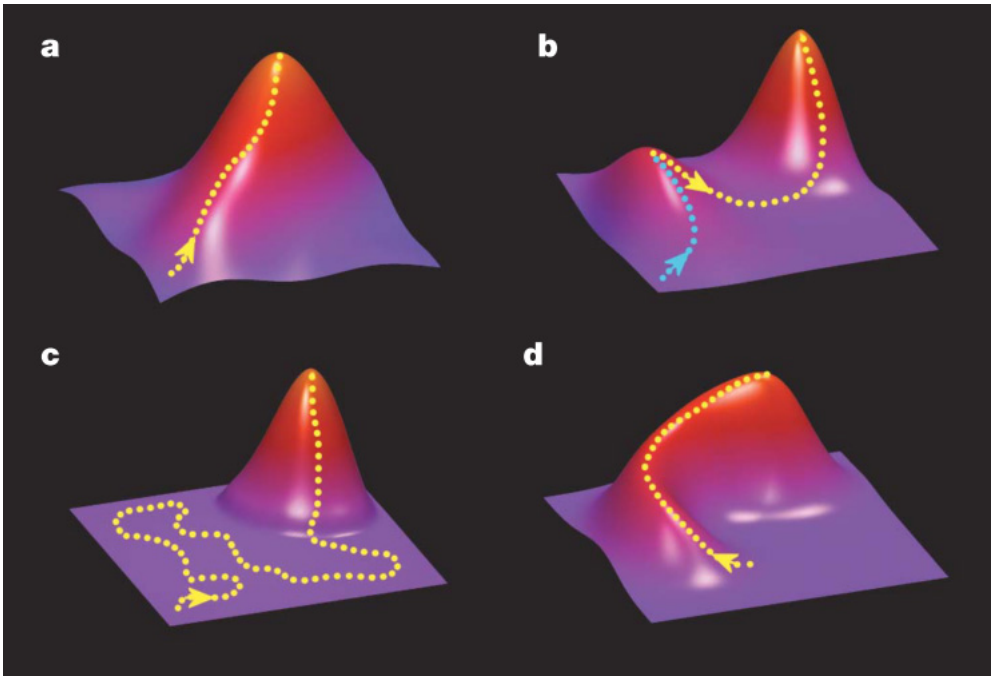


Figure 1.6: Schematic representation of fitness landscapes. a, Single smooth peak. All direct paths to the top are increasing in fitness. b, Rugged landscape with multiple peaks. The yellow path has a fitness decrease that drastically lowers its evolutionary probability. Along the blue path selection leads in the wrong direction to an evolutionary trap. c, Neutral landscape. d, Detour landscape. [Poelwijk et al., 2007].

dependant on its initial sequence [de Visser and Krug, 2014, Lobkovsky et al., 2011, Weinreich et al., 2006].

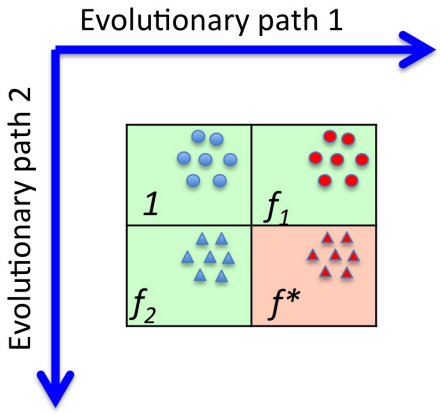


Figure 1.7: Evolutionary trajectories of two different subpopulations, each acquiring incompatible beneficial mutations ($f^*$=0) diverge in different directions.

## 1.3 Overview of the thesis

How epistatic is the functional protein space? We first addressed this issue in the paper [Povolotskaya and Kondrashov, 2010], where we investigated the rates and limits of divergence of homologous proteins in bacterial genomes. We measured the rates of evolution between pairs of proteins which shared the common ancestor as long as 3.5 billion years ago and found that even for very conserved proteins the limit of divergence was not yet reached and they are still diverging from each other. The work is presented in Chapter 2.

The most conventional explanation of this effect is the action of strong constant negative selection. But can it slow down the rate of divergence of two sequences? In order to answer this question we built a population genetics model. We found that while negative selection defines the limit of divergence, it does not drastically affect the rate with which this limit is reached [Kondrashov et al., 2010]. The work is presented in chapter 3.

The alternative explanation for the slow rate of divergence is epistatic interactions between different sites. Epistasis restricts the number of functional amino acid sequences of a protein and, thus, may slow down the rate of protein evolution. We showed that the vast majority of amino-acid positions in proteins eventually accept a substitution, but only 2% are available for evolution at any particular moment and the remaining 98% of positions cannot accept any amino-acid substitution at this moment. But when looking at the larger evolutionary scale a vast majority of sites may eventually become permitted to evolve when other, compensatory, changes occur. For those sites, which accepted a substitution, the substitution leading back to the ancestral state becomes forbidden very rapidly, indicating that selection depends on the genetic background.

Furthermore, we incorporated epistasis in theoretical model of protein evolution to capture the patterns of such slow rate of protein divergence. Based on the model we estimated average parameters of epistatic sequence space [Usmanova et al., 2014]. The work is presented in Chapter 4.

Further we looked at the patterns of single nucleotide polymorphisms in a range of metazoan species. We found that SNPs that alter protein sequence tend to occur at the same positions in different species, and the excess of such co-occurrence is much higher for species that share more recent common ancestor, indicating that patterns of standing variation are also governed by epistasis and change in the fitness effect of a certain amino acid is large enough to be seen at the polymorphic level. The work is presented in Chapter 5.

In chapter 6, we investigated the usages of different stop codon relative to bacterial genomic GC-content. We have shown that TAG stop codon is the least fit stop codon in a biologically plausible range of mutation rates and selection coefficients, and that selection acting on TAG is context dependent, increasing with the genomic GC-content.

Finally, chapter 7 is a result of international collaboration for deciphering the genome of a ctenophore *Pleurobrachia bachei*. I contributed to the project by assembling and analyses of the genome.

# 2. Sequence space and the ongoing expansion of the protein universe

Sequence space and the ongoing expansion of the protein universe.
Inna S Povolotskaya and Fyodor A Kondrashov.
Nature, 465(7300): 922-6, June 2010.
doi: 10.1038/nature09105.

Povolotskaya IS, Kondrashov FA. Sequence space and the ongoing expansion of the protein universe. Nature. 2010 Jun 17;465(7300):922-6. doi:10.1038/nature09105.

# 3. Rate of sequence divergence under constant selection.

Rate of sequence divergence under constant selection.

Alexey S Kondrashov, Inna S Povolotskaya, Dmitry N Ivankov, Fyodor A Kondrashov.

Biology direct, 5:5, January 2010.

doi: 10.1186/1745-6150-5-5.

Kondrashov AS, Povolotskaya IS, Ivankov DN, Kondrashov FA. Rate of sequence divergence under constant selection. Biol Direct. 2010 Jan 21;5:5. doi: 10.1186/1745-6150-5-5.

# 4. A model of substitution trajectories in sequence space and long-term protein evolution

A Model of Substitution Trajectories in Sequence Space and Long-Term Protein Evolution.

Dinara R Usmanova, Luca Ferretti, Inna S Povolotskaya, Peter K Vlasov, Fyodor A Kondrashov.

Molecular biology and evolution, November 2014.

doi: 10.1093/molbev/msu318.

Usmanova DR, Ferretti L, Povolotskaya IS, Vlasov PK, Kondrashov FA. A model of substitution trajectories in sequence space and long-term protein evolution. Mol Biol Evol. 2015 Feb;32(2):542-54. doi: 10.1093/molbev/msu318.

# 5. A changing fitness landscape contributes to more shared polymorphisms between closely related species.

**Inna S Povolotskaya**[1,2], **Fyodor A Kondrashov**[1,2,3]

[1]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG) 88 Dr. Aiguader, 08003 Barcelona, Spain

[2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.

[3]Institucio Catalana de Recerca i Estudis Avancats (ICREA), 23 Pg. Lluis Companys, 08010 Barcelona, Spain.

## 5.1 Introduction

"As all the species of the same genus are supposed to be descended from a common progenitor, it might be expected that they occasionally vary in an analogous manner" [Darwin, 1859]. Darwin's hypothesis has been elaborated by Vavilov as a Law of Homologous Series of Variation [Vavilov, 1922]. In Vavilov's words, "Species and genera that are genetically closely related are characterised by similar series of heritable variations with such regularity that knowing the series of form within the limits of one species, we can predict the occurrence of parallel forms in other species and genera. The more closely related the species...in the general system, the more resemblance will there be in the series of variations."

What could be the molecular mechanism of Vavilov's Law of Homologous Series? There can be three different causes to the observation that closely related species vary in a similar manner. First, they could share a polymorphism which was present in their common ancestral species and has not been neither fixated or lost. In such a scenario, shared synonymous polymorphisms are expected to be lost very rapidly, almost completely between humans and chimpanzee [Clark, 1997, Gao et al., 2014]. Nonsynonymous polymorphisms may, in theory, be maintained by balancing selection [Charlesworth, 2006], but the number of such instances is very modest [Leffler et al., 2013].

Second, polymorphisms in mutational hotspots could arise independently in different species. Under this assumption, synonymous sites have to follow Vavilov's law to a higher degree than nonsynonymous sites as synonymous sites are expected to be found in a more conserved and more slowly changing mutational context of nonsynonymous sites, thus maintaining a similar rate of mutation for a longer period of time than nonsynonymous sites, which are often flanked at least on one side with a synonymous site.

And third, similar selection pressure could favour similar polymorphisms in more related species either because they adapt to a similar environment

or because their genome occupies an adjacent position on the fitness landscape.

Recent advance of next-generation sequencing technologies and increasing availability of the variation data for a large number of genomes across different kingdoms of life have allowed to study the mechanisms of Vavilov's Law at the genomic level.

## 5.2 Results

We obtained recently published variation data for 76 non-model animal species [Romiguier et al., 2014], widely sequenced mitochondrial genomes from 53 mammalian species from GenBank [Benson et al., 2013], additional mitochondrial genomes from 5 great apes [Prado-Martinez et al., 2013] and genome-wide variation data for 16 nuclear genomes of vertebrates from Ensembl [Cunningham et al., 2014]. Using this data, we studied the probability of observing polymorphisms at both orthologous sites between two species as a function of genetic distance between those species. We focus on the relative contributions of two different factors to shared polymorphisms in different species: similarity of mutation rates and the similarity of the fitness landscape. We studied separately four-fold nonsynonymous sites and four-fold synonymous sites, using the four-fold synonymous sites as a proxy for neutral evolution.

Given the densities of polymorphisms in each species ($d_1$ and $d_2$) the expected probability that a certain site is polymorphic in both species is $d_1 * d_2$, or $E$, the expected density of sites that are polymorphic in both species. We introduce $p = O/E$, where $O$ is the observed density of two simultaneously polymorphic orthologous sites. If probability that a site carries a polymorphism is independent in the two species and selection pressure is homogenous we expect $p$ to fluctuate around 1.

Departure from $p = 1$ could be due to similarity in the rate of mutation across orthologous sites which affects both synonymous ($p_S$) and nonsynony-

mous sites ($p_N$). In addition, in nonsynonymous sites, $p_N$ measure may also be influenced by non homogeneity of selection pressure across sites in two different ways. First, strong selection may prevent the occurrence of polymorphisms in highly conserved sites, which would led to the constant inflation of $p_N$ over 1, or selection pressure might be affected by epistatic interactions between different loci and be more similar between more closely related species, due to their proximity in fitness landscape. Under this scenario $p_N$ would be inflated more in more closely related species.

A comparison of $p_S$ and $p_N$ as a function of protein divergence between the two species ($D$) may, therefore, be used to test the contribution of the factors of mutation and selection towards shared polymorphisms between species (see Methods for details on all variables).

### 5.2.1 Variation in non-model animal species

We first focus on the dataset of 76 non-model animal species from [Romiguier et al., 2014] (Figure 5.1).

Four important trends are apparent from this dataset. First, the asymptotic value of $p_S$ is close to 1, except for the most closely related species. Thus, for species >10% divergence of amino acid sequence the probability of observing the same synonymous polymorphism is indistinguishable from the random probability. Second, the $p_N$ for nonsynonymous sites does not reach 1 even for distantly related species. The equilibrium value for $p_N$ may be different if constant negative selection is acting on a substantial fraction of nonsynonymous sites. Indeed, if fraction $f$ of the nonsynonymous sites cannot under any conditions accept a polymorphism under any conditions then the expected probability of two sites having a polymorphism in the same site is $E = \frac{d_1 * d_2}{f}$. The fraction of sites that cannot accept a polymorphism is unknown, but might be determined indirectly from equilibrium value of $p_N$. In our dataset, $f = 1/p_N = 0.5$. Third, there is a decline in both $p_N$ and $p_S$ for closely related species, which could be attributed to the mutational biases. If some sites have a higher than average rate of mutation and these hotspots

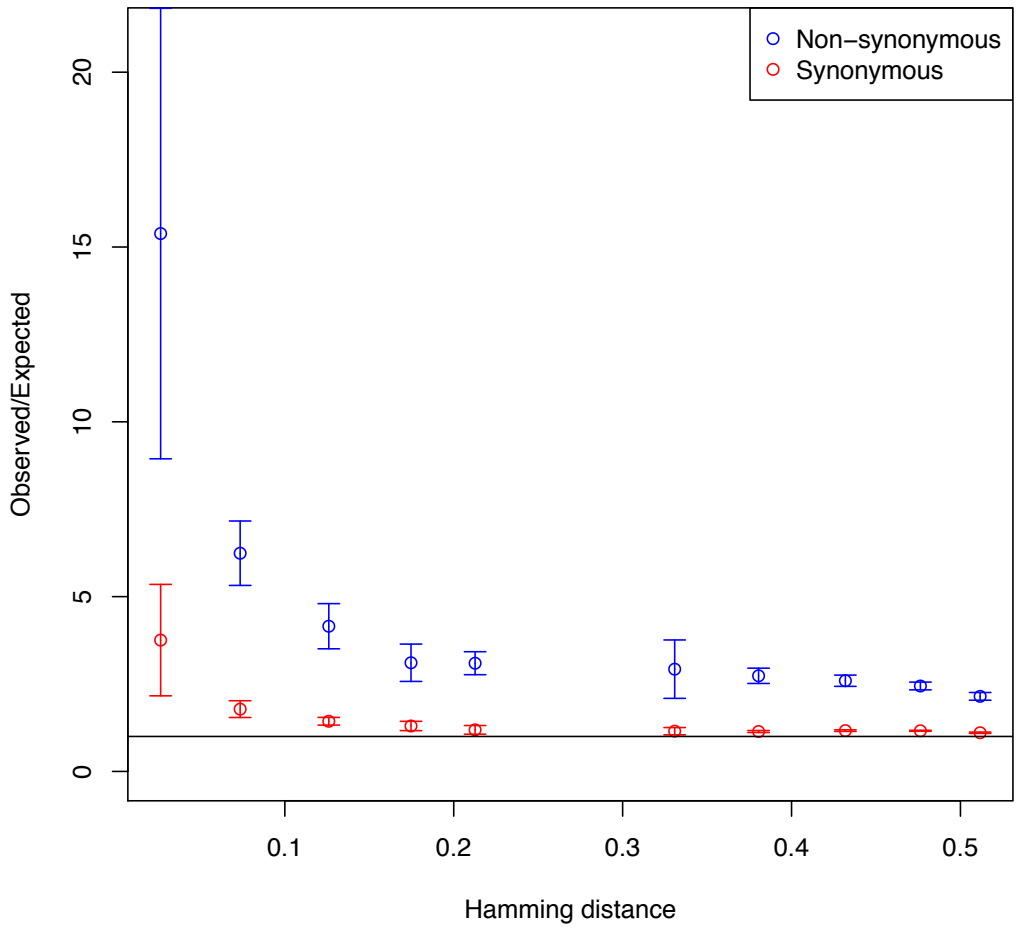


Figure 5.1: Dependence of $p_N$ and $p_S$ from protein divergence.

tend to change slowly then, when comparing the genomes of two species, some orthologous sites are more likely to be polymorphic in both species. And finally, the inflation of $p_N$ is much stronger than of $p_S$. The latter trend could potentially be explained by the combination of mutational biases and action of constant negative selection, if the ratio of $p_N/p_S$ was constant over time, but as a matter of fact it declines without reaching an equilibrium even for the largest values of D considered in our study (Figure 5.2), indicating that a shift in the fitness landscape plays a role.

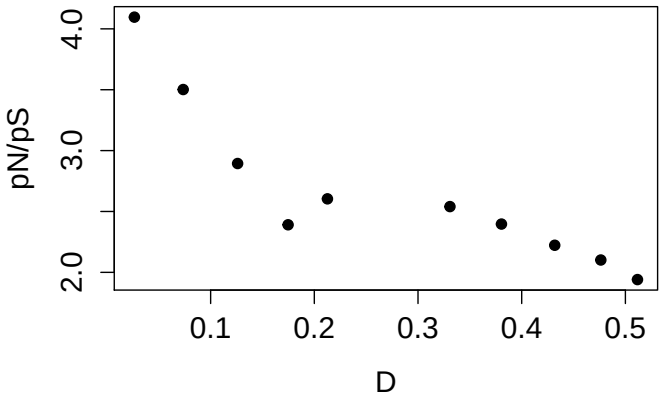


Figure 5.2: Ratio $p_N/p_S$ as a function of protein divergence.

### 5.2.2 Variation in vertebrate species

For dataset of variation in vertebrate genomes, we observe a largely congruent pattern, but with an interesting difference, $p_N$ reaches 1 (Figure 5.3).

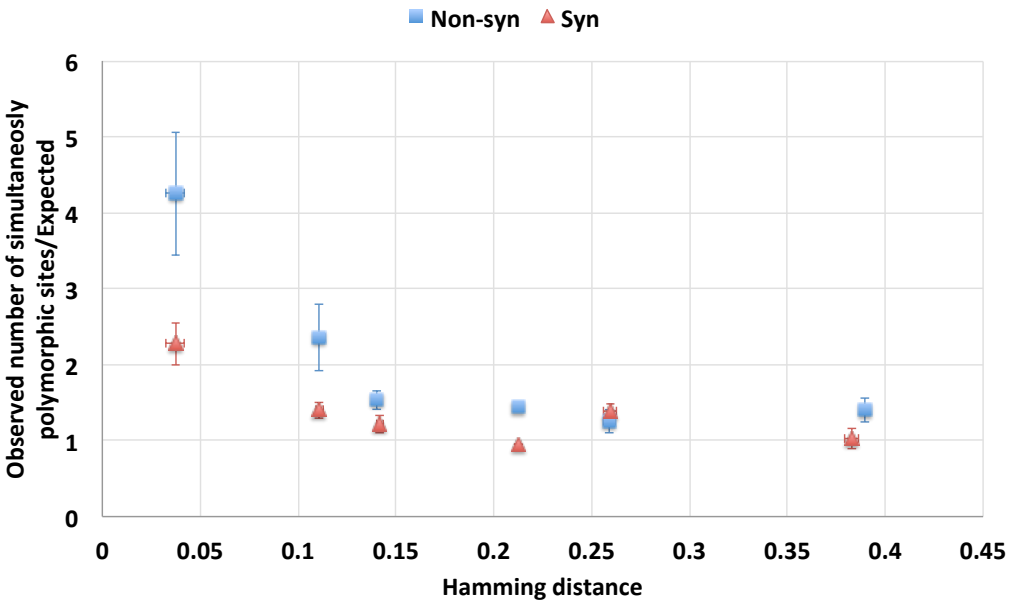


Figure 5.3: Dependence of $p_N$ and $p_S$ from protein divergence.

This dataset contains more data, as vertebrate genomes are much longer than most of the genomes in the first dataset. We were thus able to analyze patterns of variation in a more detailed fashion.

First, we analyse whether or not the specific allele states that are present as polymorphisms behave differently (Figure 5.4). There are three possibilities for the number of different allele states when considering an orthologous site that is polymorphic in two species. The site may have 2, 3 or 4 different allele states. We did not consider 4 allele states, as they happen very rarely. However, there is a big difference in the evolutionary behaviour of 2 allele (type 1) and 3 allele (type 2) orthologous polymorphic sites. The 3 allele state sites lack the time-dependence of the $O/E$ ratio in both synonymous and non-synonymous sites. In contrast, both the synonymous and nonsynonymous sites for the 2 allele state sites decline with protein distance.

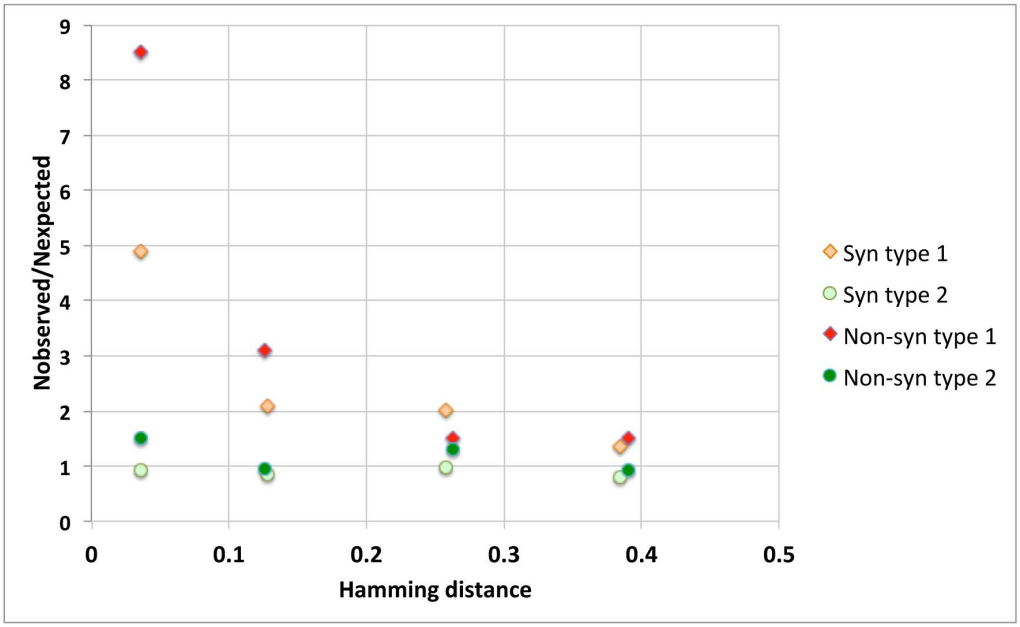


Figure 5.4: Dependence of $p_N$ and $p_S$ from protein divergence for different types of allele states.

As the form of the decay of both synonymous and nonsynonymous sites appear to be very similar for type 1 sites, both of them might have potentially been caused by a decrease in the similarity of the rate of mutation. To test this hypothesis we separated the sites into those, with an identical nucleotide context between the two considered species, and those, where the nucleotide context is different between the two. The sites where the nucleotide context is the same between the species show the time dependence for both synony-

mous and nonsynonymous sites (Figure 5.5). However, only nonsynonymous sites show the time dependence when the nucleotide context has changed between the compared species (Figure 5.6).

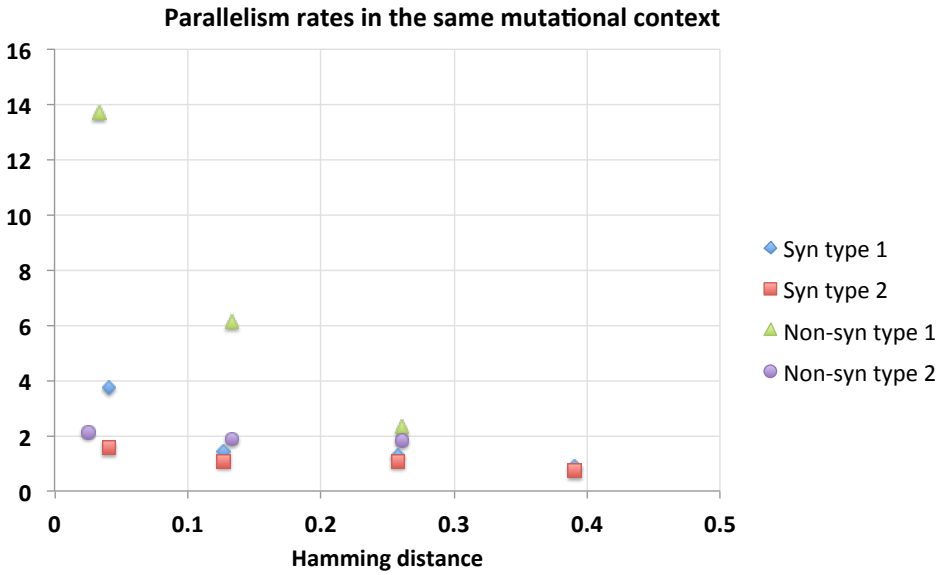


Figure 5.5: Dependence of $p_N$ and $p_S$ from protein divergence in the same mutational context.

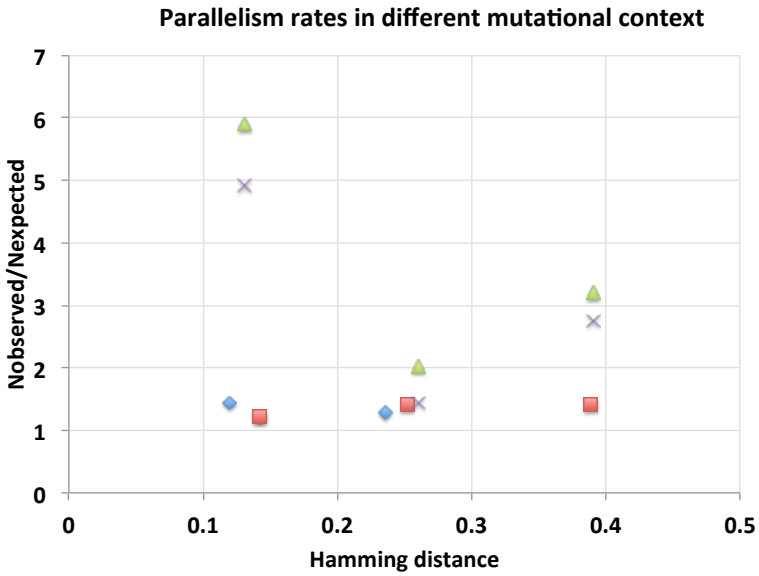


Figure 5.6: Dependence of $p_N$ and $p_S$ from protein divergence in different mutational contexts. *Blue*: 2 alleles synonymous sites; *red*: 3 alleles synonymous sites; *green*: 2 alleles nonsynonymous sites; *purple*: 3 alleles nonsynonymous sites

These data suggest, that while mutational biases cause a substantial

inflation of both $p_S$ and $p_N$ in vertebrates genomes, change in selection pressure is also in play.

### 5.2.3 Variation in mammalian mitochondrial genomes

Finally, we looked at the mitochondrial data. The mitochondrial data are the cleanest in the sense that synonymous sites very quickly reach the random expectation, while the non-synonymous sites are substantially above 1, reaching the plateau long after the synonymous sites. For mitochondrial genes we observe that the probability of synonymous sites harbouring a polymorphism in different species is indistinguishable from random regardless of the distance between the two species (Figure 5.7).

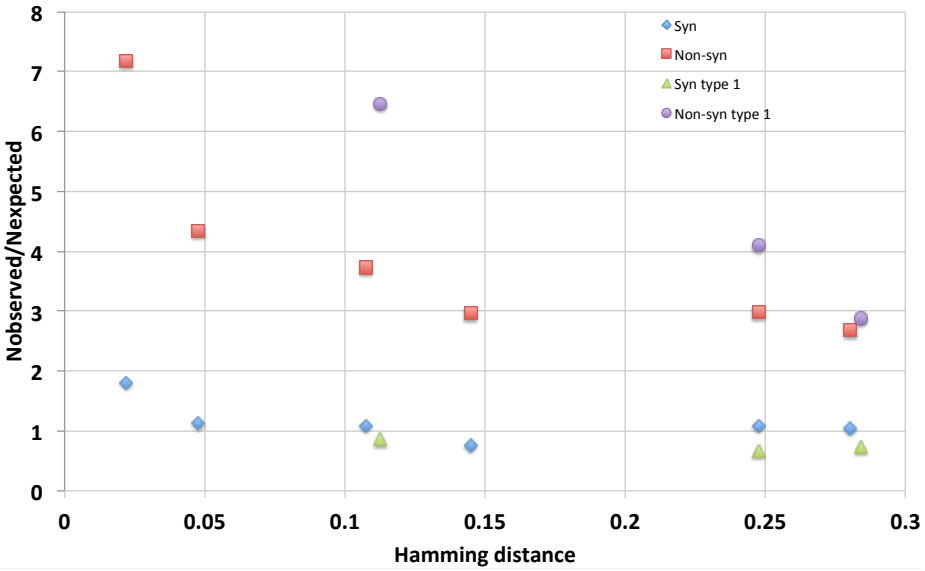


Figure 5.7: Dependence of $p_N$ and $p_S$ from protein divergence for mitochondrial genomes.

### 5.2.4 Simplified model of epistatic interactions

We further simplified model from [Usmanova et al., 2014] which takes epistatic interactions into account directly and used it to fit the data from the dataset of 76 non-model animal genomes.

Let us assume that there is $N(t)$ the number of sites which could accommodate a polymorphism in a given species and $f(t)$ the frequency of such sites. Let us further assume, that the sites which can accommodate a polymorphism lose this feature with the rate $a$ and the sites which initially could not accommodate a polymorphisms gain this feature with the rate $b$. Then,

$$\frac{df(t)}{dt} = -(a+b)f(t) + b,$$

and in equilibrium

$$f = \frac{b}{a+b}.$$

In this simplified model, the equation for the probability, that the site can accommodate a polymorphism $r(t)$ has the same form as $f(t)$ and has two solutions, based on the state in the $t = 0$. We are interested in the solution which corresponds to the situation when the site could accommodate a polymorphism in the $t = 0$, as we could consider one of the species as a point of departure and are restricted to the sites which can accommodate a polymorphism in both species. The solution for the probability not to lose the ability to accommodate a polymorphism is

$$r(t) = f + (1-f)e^{-(a+b)t}.$$

Our calculations for the expected number of simultaneously polymorphic sites can be rewritten as following:

$$N_{homogenious} = \frac{n_1}{N_{total}} \frac{n_2}{N_{total}} N_{total}$$

And under the epistatic model:

$$N_{epistatic} = \frac{n_1}{N_{total}f} \frac{n_2}{N_{total}f} N_{total} f r(t)$$

The ratio $\frac{N_{epistatic}}{N_{homogenous}}$ is then $p_N$ from our previous calculations and is

$$p_N(t) = \frac{1}{f} r(t) = 1 + \frac{a}{b} e^{-(a+b)t}$$

We thus fit our data with the exponential function (Figure 5.8) and get an approximation of $p_N(t) \approx 0.12 + 0.88e^{-2.6d}$.

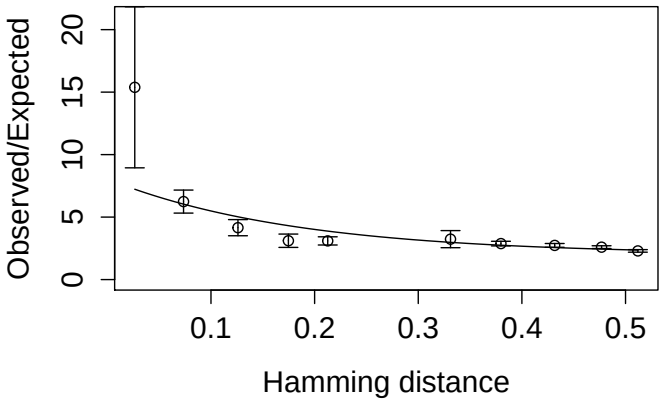


Figure 5.8: Exponential fit of the data

Based on the approximation we can calculate the proportion of sites which are allowed for accommodating a polymorphism in orthologous sites of two species as a function of protein divergence between these species (Figure 5.9).

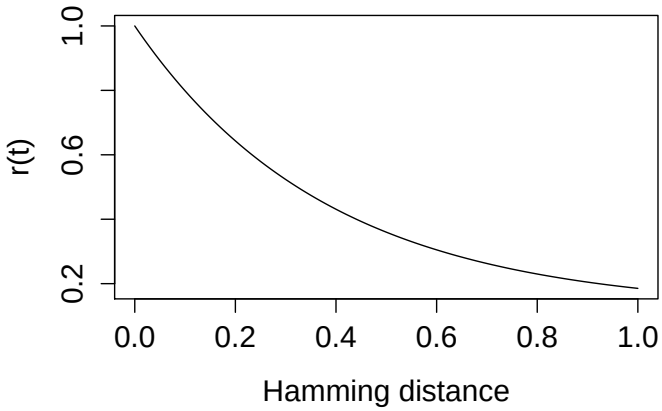


Figure 5.9: Proportion of sites which are allowed for accommodating a polymorphism in orthologous sites of two species as a function of protein divergence between these species

## 5.3 Discussion

Here we study the possible molecular mechanisms of the Vavilov's Law of Homologous series. We find that there is persistent increase of the number of nonsynonymous sites which are polymorphic in two closely related species in all our datasets, animals, vertebrates and mitochondrial genomes, indicating that this pattern is caused by a basic principle underlying protein evolution rather than specific to a particular clade. We further show, that both mutational biases and a similarity in the fitness landscape led to increase of a number of polymorphisms in orthologous sites in closely related species.

In principle, for very closely related species, polymorphisms could be shared due to the origin in the shared ancestral species, but the number of neutral ancestral polymorphisms which are expected to be found in two even closely related species is very low, on the scale from 0 to 5 [Clark, 1997], while for the most closely related species we observe on average $1300 \pm 350$ synonymous sites which are polymorphic in two species and $350 \pm 100$ nonsynonymous sites. Trans-species nonsynonymous polymorphisms could theoretically be maintained by balancing section, but again, the number of such instances reported is very few (2 between humans and chimpanzees) [Charlesworth, 2006, Key et al., 2014, Leffler et al., 2013] and is unlikely to explain the inflation of neither $p_S$ or $p_N$.

Excess of parallel polymorphisms in synonymous sites indicates that a fraction of parallel nonsynonymous polymorphisms in closely related species is caused by heterogeneity of mutational rates [Hwang and Green, 2004]. Thus, a portion of similar phenotypic variation in traits in closely related species could be attributed to neutral-like processes, and for the protein divergence of 3% we expect at least twice as many neutral nonsynonymous polymorphisms to segregate in two species than for more distant species with protein divergence of 50% due to mutational biases alone.

Finally, the decline of the density of parallel nonsynoymous changes might be caused by the shift in the fitness landscape. Our data suggest, that in most

animal genomes 50% of nonsynonymous sites are under strong negative selection and can not accommodate a polymorphism. The rest of the sites show a decrease of the rate of parallel variation with protein divergence, which indicates a change in the selection pressure acting at the orthologous sites. There are two likely phenomenons which cause a change in the selection pressure. First, more closely related species need to adapt to more similar environments. However, positive selection is believed to be a rare event and fixation of the selectively advantageous polymorphisms happens faster than both neutral and deleterious [Haldane, 2008, Kimura, 1957] . Recent reports [Breen et al., 2012, Povolotskaya and Kondrashov, 2010] assert ubiquity of genetic interactions between different mutations. Our data are in high concordance with epistatic mode of evolution, further elaborating the hypothesis, that the fitness of a certain allele depends on the genetic background. Moreover, in order to see a change in the selection pressure associated with a polymorphism its selection coefficient must experience a dramatic change, effectively switching from evolvable to evolutionary constant state.

Convergent evolution between closely related species is known to be widespread [Bazykin et al., 2007, Rokas and Carroll, 2008] and decline with time[Naumenko et al., 2012, Povolotskaya and Kondrashov, 2010, Rogozin et al., 2008]. Here we show, that the decline of the fitness happens rapidly, and the fitness effect of a particular variant will be different in 20% of sites for the species with only 10% of median protein divergence (which roughly corresponds to protein divergence between human and mouse) and in 65% of the sites for the species with 40% of protein divergence (about divergence between human and fishes).

## 5.4 Methods

We obtained available sequences of complete mammalian mitochondrial genomes from Genbank [Benson et al., 2013], additional mitochondrial

genomes from 5 primate species [Prado-Martinez et al., 2013] and vertebrate nuclear genomes with available variation data from Ensembl [Cunningham et al., 2014]. Furthermore, we utilized the dataset ecently published variation data for 76 non-model animal species [Romiguier et al., 2014], that contains the genetic variability species from different phylogenetic groups.

### 5.4.1 Data preparation

For the dataset of vertebrates nuclear genomes, we extracted one-to-one orthologous clusters based on MethaPhOrs predictions [Pryszcz et al., 2011] and aligned each set of orthologs using T-Coffee[Notredame et al., 2000]. We further removed inaccurate parts of alignments using trimAl [Capella-Gutiérrez et al., 2009] standard gappyout method. We further excluded CpG sites from the analysis, as a known source of strong mutational biases [Bird, 1980]. For the dataset of mammalian mitochondrial genomes, we only used genomes which encode for exactly 13 common mitochondrial protein-coding genes, thus excluding orthology prediction step. We aligned each set of orthologs using T-Coffee [Notredame et al., 2000]. For the dataset of non-model metazoan genomes, we identified orthologs using reciprocal blast hit approach [Altschul et al., 1990, Tatusov et al., 1997]. We further removed 10% of the predicted orthologous pairs with the lowest identity, in order to decrease the number of falsely identified orthology. We created pairwise protein alignments for all the orthologous pairs between every pair of species using MUSCLE [Edgar, 2004] and further trimmed gaps flanking regions of length 5 amino acid.

Protein divergence between two species was calculated as a median pairwise protein divergence in the aligned regions of the proteins.

### 5.4.2 Calculation of expected probabilities

For every species we then identify biallelic four-fold synonymous and four-fold nonsynonymous single nucleotide polymorphisms and their corresponding

densities ($d_S$ and $d_N$) as a number of polymorphic sites of each type divided by the total number of sites of this type in the pairwise alignment (our-fold synonymous and four-fold nonsynonymous respectively).

The expected under homogenous neutral model density of sites ($E$), polymorphic in two species is simply the product of the densities of polymorphisms in these species, such as $E_S = d_S^1 * d_S^2$ for four-fold synonymous sites and $E_N = d_N^1 * d_N^2$ for four-fold nonsynonymous sites. For pairs of species with expected total number of simultaneously polymorphic sites >= 5 we took the ratio of observed ($O$) to expected density of sites that were polymorphic in both species ($p_N = E_N/O_N$, $p_S = E_S/O_S$).

### 5.4.3 Different categories of allelic states

There are three different categories of allelic states between two orthologous biallelic polymorphic sites. First, the same two alleles could be segregating in both species (type 1), species may share one allele (type 2) and 0 alleles (type 3). We found that the total number of sites which belong to type 3 category to be very rare and excluded such sites from the analysis.

In order to estimate the densities of polymorphisms of each type for every species we then reconstructed the ancestral states of each site using maximum parsimony approach and used the closest species in our dataset as an outgroup (we then excluded this pair from the subsequent analysis). The cites, where the closest species was either polymorphic or different from either segregating allele, were excluded from the analysis. Then the densities of polymorphisms were calculated, as a number of polymorphic sites of a certain type (type 1 or 2, both for four-fold synonymous and four-fold nonsynonymous sites) divided by the number of sites where such polymorphism could arise, based on the ancestral sequence. Subsequent analysis follows the procedure described above.

### 5.4.4 Different mutational context

We investigated the influence of mutational biases based on the conservation of nucleotide context of the polymorphism between pairs of species. If nucleotides both upstream and downstream from the polymorphic site were identical, we assigned these pair of polymorphisms to segregating in the same mutational context, while if at least one of the nucleotide either upstream or downstream was different, we assigned the pair of polymorphisms to segregating in different mutational contexts.

# 6. Context dependent selection acting on stop codons in bacteria

Bacterial organisms differ dramatically in their use of different nucleotides, the phenomenon which is known as GC-content bias: the use of G and C varies from less than 20% in some species, to more than 70% in others. This bias is even stronger in the third positions of codons of the protein coding regions, where usage of GC-terminating codons increases with genomic GC-content, and could be as low as 3% or as high as 99%. Interestingly, frequencies of stop codon do not follow this pattern for all stop codons: while TAA and TGA stop codon usage is correlated with genomic GC-content, the usage of TAG is uniform. Here, we develop a simple model, which captures the frequencies of stop codon across genomes with different GC-contents and show that the selection, acting on the TAG stop codon must be dependant on the genomic GC-content, and this codon is universally associated with lower fitness.

Stop codons in bacteria are not selectively equivalent.

Inna S Povolotskaya, Fyodor A Kondrashov, Alice Ledda, Peter K Vlasov.

Biology direct, 7(1): 30, January 2012.

doi: 10.1186/1745-6150-7-30.

Povolotskaya IS, Kondrashov FA, Ledda A, Vlasov PK. Stop codons in bacteria are not selectively equivalent. Biol Direct. 2012 Sep 13;7:30. doi: 10.1186/1745-6150-7-30.
http://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-7-30

# 7. The ctenophore genome and the evolutionary origins of neural systems.

The ctenophore genome and the evolutionary origins of neural systems.

Leonid L. Moroz, Kevin M. Kocot, Mathew R. Citarella, Sohn Dosung, Tigran P. Norekian, Inna S. Povolotskaya, Anastasia P. Grigorenko, Christopher Dailey, Eugene Berezikov, Katherine M. Buckley, Andrey Ptitsyn, Denis Reshetov, Krishanu Mukherjee, Tatiana P. Moroz, Yelena Bobkova, Fahong Yu, Vladimir V. Kapitonov, Jerzy Jurka, Yuri V. Bobkov, Joshua J. Swore, David O. Girardo, Alexander Fodor, Fedor Gusev, Rachel Sanford, Rebecca Bruders, Ellen Kittler, Claudia E. Mills, Jonathan P. Rast, Romain Derelle, Victor V. Solovyev, Fyodor A. Kondrashov, Billie J. Swalla, Jonathan V. Sweedler, Evgeny I. Rogaev, Kenneth M. Halanych, Andrea B. Kohn.

Nature 510, 109-114 (05 June 2014).

doi:10.1038/nature13400.

Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, Ptitsyn A, Reshetov D, Mukherjee K, Moroz TP, Bobkova Y, Yu F, Kapitonov VV, Jurka J, Bobkov YV, Swore JJ, Girardo DO, Fodor A, Gusev F, Sanford R, Bruders R, Kittler E, Mills CE, Rast JP, Derelle R, Solovyev VV, Kondrashov FA, Swalla BJ, Sweedler JV, Rogaev EI, Halanych KM, Kohn AB. The ctenophore genome and the evolutionary origins of neural systems. Nature. 2014 Jun 5;510(7503):109-14. doi: 10.1038/nature13400.

# 8. Conclusions

- Ancient homologous proteins which shared common ancestor 3.5 billion years ago have not yet reached the limits of sequence divergence defined by the functional constraints.
- Constant selection framework predicts equilibrium limits of protein divergence to be lower than the observed values. In turn, rate of approaching equilibrium is predicted to be of the same order as this rate at the synonymous sites.
- The rate of convergent evolution declines with protein distance, indicating rapid turnover of fitness values associated with individual amino acids.
- Model which directly incorporates epistatic interactions between amino acids accurately captures the observed values of the rates of convergent evolution. Parameters of epistasis predicted by the model differ substantially between different protein families.
- Mutational biases and a similarity in the fitness landscape led to increase of a number of parallel polymorphisms in orthologous sites in closely related species.
- In bacterial species, TAG stop codon is universally associated with lower fitness. Selection acting on TAG stop codon is dependant on the genomic GC-content and is stronger in GC-reach genomes.
- The genome of *Pleurobrachia bachei* was assembled. Analyses of the genome demonstrated that ctenophore neural systems, likely, evolved independently from those in other animals.

# A. List of publications

1. Kondrashov AS, **Povolotskaya IS**, Ivankov DN, Kondrashov FA. Rate of sequence divergence under constant selection. *Biology Direct. 2010 Jan 21.*

2. **Povolotskaya IS**, Kondrashov FA. Sequence space and the ongoing expansion of the protein universe. *Nature. 2010 Jun 17.*

3. **Povolotskaya IS**, Kondrashov FA, Ledda A, Vlasov PK. Stop codons in bacteria are not selectively equivalent. *Biology Direct. 2012 Sep 13.*

4. Moroz LL, Kocot KM, Citarella MR, Dosung S, **Povolotskaya IS**, Norekian TP, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, Ptitsyn A, Reshetov D, Mukherjee K, Moroz TP, Bobkova Y, Yu F, Kapitonov VV, Jurka J, Bobkov YV, Swore JJ, Girardo DO, Fodor A, Gusev F, Sanford R, Bruders R, Kittler E, Mills CE, Rast JP, Derelle R, Solovyev VV, Kondrashov FA, Swalla BJ, Sweedler JV, Rogaev EI, Halanych KM, Kohn AB. The Ctenophore Genome and the Evolutionary Origins of Neural Systems. *Nature. 2014 May 21.*

5. Lasky JR, Des Marais DL, Lowry DB, **Povolotskaya IS**, McKay JK, Richards JH, Keitt TH, Juenger TE. Natural variation in abiotic stress responsive gene expression and local adaptation to climate in Arabidopsis thaliana. *Mol Biol Evol. 2014 May 21.*

6. Usmanova DR, Ferretti L, **Povolotskaya IS**, Vlasov PK, Kondrashov FA. A model of long-term protein evolution. *Mol Biol Evol, 2014 Nov 17*.

7. **Povolotskaya IS**, Kondrashov FA. A changing fitness landscape contributes to more shared polymorphisms between closely related species.(In preparation).

# Bibliography

S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10, October 1990.

Claudia Bank, Ryan T Hietpas, Jeffrey D Jensen, and Daniel N A Bolon. A systematic survey of an intragenic epistatic landscape. *Molecular biology and evolution*, pages 1–29, November 2014.

Anja Baresić, Lisa E M Hopcroft, Hubert H Rogers, Jacob M Hurst, and Andrew C R Martin. Compensated pathogenic deviations: analysis of structural effects. *Journal of molecular biology*, 396(1):19–30, February 2010.

Michael C Bassik, Martin Kampmann, Robert Jan Lebbink, Shuyi Wang, Marco Y Hein, Ina Poser, Jimena Weibezahn, Max A Horlbeck, Siyuan Chen, Matthias Mann, Anthony A Hyman, Emily M Leproust, Michael T McManus, and Jonathan S Weissman. A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. *Cell*, 152 (4):909–22, February 2013.

Georgii a Bazykin, Fyodor a Kondrashov, Michael Brudno, Alexander Poliakov, Inna Dubchak, and Alexey S Kondrashov. Extensive parallelism in protein evolution. *Biology direct*, 2:20, January 2007.

Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic acids research*, 41(Database issue):D36–42, January 2013.

A P Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research*, 8(7):1499–504, April 1980.

Sebastian Bonhoeffer, Colombe Chappey, Neil T Parkin, Jeanette M Whitcomb, and Christos J Petropoulos. Evidence for positive epistasis in HIV-1. *Science (New York, N.Y.)*, 306(5701):1547–50, November 2004.

Michael S Breen, Carsten Kemena, Peter K Vlasov, Cedric Notredame, and Fyodor a Kondrashov. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–8, October 2012.

Christina L Burch and Lin Chao. Epistasis and its relationship to canalization in the RNA virus phi 6. *Genetics*, 167(2):559–67, June 2004.

David Butcher. Muller's ratchet, epistasis and mutation effects. *Genetics*, 141(1):431–7, September 1995.

Alexandra B Byrne, Matthew T Weirauch, Victoria Wong, Martina Koeva, Scott J Dixon, Joshua M Stuart, and Peter J Roy. A global analysis of genetic interactions in Caenorhabditis elegans. *Journal of Biology*, 6(8), 2007.

Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, 25(15):1972–3, August 2009.

Deborah Charlesworth. Balancing selection and its effects on sequences in nearby genome regions. *PLoS genetics*, 2(4):e64, April 2006.

Hsin-Hung Chou, Hsuan-Chao Chiu, Nigel F Delaney, Daniel Segrè, and Christopher J Marx. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science (New York, N.Y.)*, 332(6034):1190–2, June 2011.

A G Clark. Neutral behavior of shared polymorphism. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15): 7730–4, July 1997.

H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, October 2002.

Michael Costanzo, Carolyn S Sevier, Huiming Ding, Judice L Y Koh, Kiana Toufighi, Sara Mostafavi, Franco J Vizeacoumar, Solmaz Alizadeh, Sondra Bahr, Renee L Brost, Yiqun Chen, Corey Nislow, Olga G Troyanskaya, Howard Bussey, Gary D Bader, Brenda J Andrews, and Charles Boone. The Genetic Landscape of a Cell. *Science*, 327(425), 2010.

J A Coyne. Genetics and speciation. *Nature*, 355(6360):511–5, February 1992.

Fiona Cunningham, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E Hunt, Sophie H Janacek, Nathan Johnson, Thomas Juettemann, Andreas K Kähäri, Stephen Keenan, Fergal J Martin, Thomas Maurel, William McLaren, Daniel N Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatelli, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P Wilder, Amonida Zadissa, Bronwen L Aken, Ewan Birney, Jennifer Harrow, Rhoda Kinsella, Matthieu Muffato, Magali Ruffier, Stephen M J Searle, Giulietta Spudich, Stephen J Trevanion, Andy Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2015. *Nucleic acids research*, October 2014.

Charles. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* JOHN MURRAY, LONDON, 1859.

Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the

human genome to be under selective constraint using GERP++. *PLoS computational biology*, 6(12):e1001025, January 2010.

J Arjan G M de Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature reviews. Genetics*, 15(7):480–90, July 2014.

T. Dobzhansky. *Genetics and the Origin of Species*. Columbia University Press, 1937.

Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–7, January 2004.

Santiago F Elena and Richard E Lenski. Test of synergistic interactions among deleterious mutations in bacteria. *Nature*, 390(6658):395–8, November 1997.

Santiago F Elena, Ricard V Solé, and Josep Sardanyés. Simple genomes, complex interactions: epistasis in RNA virus. *Chaos (Woodbury, N.Y.)*, 20 (2):026106, June 2010.

Suzanne Estes, Patrick C Phillips, and Dee R Denver. Fitness recovery and compensatory evolution in natural mutant lines of C. elegans. *Evolution; international journal of organic evolution*, 65(8):2335–44, August 2011.

Ziyue Gao, Molly Przeworski, and Guy Sella. Footprints of ancient balanced polymorphisms in genetic variation data from closely related species. *Evolution*, pages 1–42, November 2014.

Lizhi Ian Gong, Marc a Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, 2:e00631, January 2013.

J. B. S. Haldane. A mathematical theory of natural and artificial selection. (Part VI, Isolation.). *Mathematical Proceedings of the Cambridge Philosophical Society*, 26(02):220, October 2008.

Ryan Hietpas, Benjamin Roscoe, Li Jiang, and Daniel N a Bolon. Fitness analyses of all possible point mutations for regions of genes in yeast. *Nature protocols*, 7(7):1382–96, July 2012.

Thomas Horn, Thomas Sandmann, Bernd Fischer, Elin Axelsson, Wolfgang Huber, and Michael Boutros. Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nature methods*, 8(4): 341–6, April 2011.

Dick G Hwang and Phil Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39):13994–4001, September 2004.

Felix M Key, João C Teixeira, Cesare de Filippo, and Aida M Andrés. Advantageous diversity maintained by balancing selection in humans. *Current opinion in genetics & development*, 29C:45–51, August 2014.

Aisha I Khan, Duy M Dinh, Dominique Schneider, Richard E Lenski, and Tim F Cooper. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science (New York, N.Y.)*, 332(6034):1193–6, June 2011.

M Kimura. Some problems of stochastic processes in genetics. *The Annals of Mathematical Statistics*, 1957.

M Kimura. The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*, 84, 1985.

A S Kondrashov and J F Crow. Haploidy or diploidy: which is better? *Nature*, 351(6324):314–5, May 1991.

A S Kondrashov and Fyodor A Kondrashov. Interactions among quantitative traits in the course of sympatric speciation. *Nature*, 400(6742):351–4, July 1999.

Alexey S Kondrashov. Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336, 1988.

Alexey S Kondrashov, Shamil Sunyaev, and Fyodor A Kondrashov. Dobzhansky-Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 99 (23):14878–83, November 2002.

Alexey S Kondrashov, Inna S Povolotskaya, Dmitry N Ivankov, and Fyodor a Kondrashov. Rate of sequence divergence under constant selection. *Biology direct*, 5:5, January 2010.

Sergey Kryazhimskiy, Jeremy a Draghi, and Joshua B Plotkin. Evolution. In evolution, the sum is less than its parts. *Science (New York, N.Y.)*, 332 (6034):1160–1, June 2011.

Ellen M Leffler, Ziyue Gao, Susanne Pfeifer, Laure Ségurel, Adam Auton, Oliver Venn, Rory Bowden, Ronald Bontrop, Jeffrey D Wall, Guy Sella, Peter Donnelly, Gilean McVean, and Molly Przeworski. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science (New York, N.Y.)*, 339(6127):1578–82, March 2013.

Ben Lehner, Catriona Crombie, Julia Tischler, Angelo Fortunato, and Andrew G Fraser. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nature genetics*, 38(8):896–903, August 2006.

Alexander E Lobkovsky, Yuri I Wolf, and Eugene V Koonin. Predictability of evolutionary trajectories in fitness landscapes. *PLoS computational biology*, 7(12):e1002302, December 2011.

Marta Luksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507(7490):57–61, March 2014.

Mark Lunzer, G Brian Golding, and Antony M Dean. Pervasive cryptic epistasis in molecular evolution. *PLoS genetics*, 6(10):e1001162, October 2010.

Sophie Maisnier-Patin, John R Roth, Asa Fredriksson, Thomas Nyström, Otto G Berg, and Dan I Andersson. Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nature genetics*, 37(12):1376–9, December 2005.

Teri a Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia a Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven a McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, October 2009.

J Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225(5232):563–4, February 1970.

David M McCandlish, Etienne Rajon, Premal Shah, Yang Ding, and Joshua B Plotkin. The role of epistasis in protein evolution. *Nature*, 497(7451):E1–2; discussion E2–3, May 2013.

Margarita V Meer, Alexey S Kondrashov, Yael Artzy-Randrup, and Fyodor a Kondrashov. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature*, 464(7286):279–82, March 2010.

Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, and Stanley Fields. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA (New York, N.Y.)*, 19(11):1537–51, November 2013.

Sergey a Naumenko, Alexey S Kondrashov, and Georgii a Bazykin. Fitness conferred by replaced amino acids declines with time. *Biology letters*, 8(5): 825–8, October 2012.

C Notredame, D G Higgins, and J Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17, September 2000.

H A Orr. The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics*, 139(4):1805–13, April 1995.

Patrick C Phillips. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics*, 9 (11):855–67, November 2008.

Frank J Poelwijk, Daniel J Kiviet, Daniel M Weinreich, and Sander J Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–6, January 2007.

Frank J Poelwijk, Sorin Tănase-Nicola, Daniel J Kiviet, and Sander J Tans. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of theoretical biology*, 272(1):141–4, March 2011.

Inna S Povolotskaya and Fyodor A Kondrashov. Sequence space and the ongoing expansion of the protein universe. *Nature*, 465(7300):922–6, June 2010.

Javier Prado-Martinez, Peter H Sudmant, Jeffrey M Kidd, Heng Li, Joanna L Kelley, Belen Lorente-Galdos, Krishna R Veeramah, August E Woerner, Timothy D O'Connor, Gabriel Santpere, Alexander Cagan, Christoph Theunert, Ferran Casals, Hafid Laayouni, Kasper Munch, Asger Hobolth, Anders E Halager, Maika Malig, Jessica Hernandez-Rodriguez, Irene Hernando-Herraez, Kay Prüfer, Marc Pybus, Laurel Johnstone, Michael Lachmann, Can Alkan, Dorina Twigg, Natalia Petit, Carl Baker, Fereydoun Hormozdiari, Marcos Fernandez-Callejo, Marc Dabad, Michael L Wilson,

Laurie Stevison, Cristina Camprubí, Tiago Carvalho, Aurora Ruiz-Herrera, Laura Vives, Marta Mele, Teresa Abello, Ivanela Kondova, Ronald E Bontrop, Anne Pusey, Felix Lankester, John A Kiyang, Richard A Bergl, Elizabeth Lonsdorf, Simon Myers, Mario Ventura, Pascal Gagneux, David Comas, Hans Siegismund, Julie Blanc, Lidia Agueda-Calpena, Marta Gut, Lucinda Fulton, Sarah A Tishkoff, James C Mullikin, Richard K Wilson, Ivo G Gut, Mary Katherine Gonder, Oliver A Ryder, Beatrice H Hahn, Arcadi Navarro, Joshua M Akey, Jaume Bertranpetit, David Reich, Thomas Mailund, Mikkel H Schierup, Christina Hvilsom, Aida M Andrés, Jeffrey D Wall, Carlos D Bustamante, Michael F Hammer, Evan E Eichler, and Tomas Marques-Bonet. Great ape genetic diversity and population history. *Nature*, 499(7459):471–5, July 2013.

Leszek P Pryszcz, Jaime Huerta-Cepas, and Toni Gabaldón. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic acids research*, 39 (5):e32, March 2011.

Igor B Rogozin, Karen Thomson, Miklós Csürös, Liran Carmel, and Eugene V Koonin. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. *Biology direct*, 3:7, January 2008.

Antonis Rokas and Sean B Carroll. Frequent and widespread parallel evolution of protein sequences. *Molecular biology and evolution*, 25(9):1943–53, September 2008.

J. Romiguier, P. Gayral, M. Ballenghien, a. Bernard, V. Cahais, a. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, E. Loire, J. M. Lourenco, B. Nabholz, C. Roux, G. Tsagkogeorga, a. a. T. Weber, L. a. Weinert, K. Belkhir, N. Bierne, S. Glémin, and N. Galtier. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, August 2014.

Rafael Sanjuán and Santiago F Elena. Epistasis correlates to genomic complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 103(39):14402–5, September 2006.

Rafael Sanjuán, Andrés Moya, and Santiago F Elena. The contribution of epistasis to the architecture of fitness in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43): 15376–9, October 2004.

Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard a Gibbs, W James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–50, August 2005.

Onuralp Soylemez and Fyodor A Kondrashov. Estimating the Rate of Irreversibility in Protein Evolution. *Genome Biology and Evolution*, 4(12): 1213–1222, 2012.

Shamil Sunyaev, Vasily Ramensky, Ina Koch, W Lathe, Alexey S Kondrashov, and Peer Bork. Prediction of deleterious human alleles. *Human molecular genetics*, 10(6):591–7, March 2001.

Béla Szamecz, Gábor Boross, Dorottya Kalapis, Károly Kovács, Gergely Fekete, Zoltán Farkas, Viktória Lázár, Mónika Hrtyan, Patrick Kemmeren, Marian J a Groot Koerkamp, Edit Rutkai, Frank C P Holstege, Balázs Papp, and Csaba Pál. The Genomic Landscape of Compensatory Evolution. *PLoS biology*, 12(8):e1001935, August 2014.

R L Tatusov, E V Koonin, and D J Lipman. A genomic perspective on protein families. *Science (New York, N.Y.)*, 278(5338):631–7, October 1997.

Olivier Tenaillon, Alejandra Rodríguez-Verdugo, Rebecca L Gaut, Pamela McDonald, Albert F Bennett, Anthony D Long, and Brandon S Gaut. The

molecular diversity of adaptive convergence. *Science (New York, N.Y.)*, 335(6067):457–61, January 2012.

Sandra Trindade, Ana Sousa, Karina Bivar Xavier, Francisco Dionisio, Miguel Godinho Ferreira, and Isabel Gordo. Positive epistasis drives the acquisition of multidrug resistance. *PLoS genetics*, 5(7):e1000578, July 2009.

Dinara R Usmanova, Luca Ferretti, Inna S Povolotskaya, Peter K Vlasov, and Fyodor a Kondrashov. A model of substitution trajectories in sequence space and long-term protein evolution. *Molecular biology and evolution*, November 2014.

Nikolay I. Vavilov. The law of homologous series in variation. *Journal of Genetics*, 12:47–89, 1922.

Yinhua Wang, Carolina Díaz Arenas, Daniel M Stoebel, and Tim F Cooper. Genetic background affects epistatic interactions between two beneficial mutations. *Biology letters*, 9(1):20120328, February 2013.

Daniel M Weinreich, Nigel F Delaney, Mark a Depristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science (New York, N.Y.)*, 312(5770):111–4, April 2006.

Sewall Wright. The reoles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of The Sixth International Congress of Genetics*, 1:356–366, 1932.

Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–8, January 2012.