Genetic diversity and geographic patterns of human herpesvirus 4 and 6

Marco Telford

DOCTORAL THESIS UPF / 2017

THESIS DIRECTORS:

Dr. Arcadi Navarro i Cuartiellas

Dr. Gabriel Santpere Baró

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

UNIVERSITAT POMPEU FABRA

Abstract

This thesis focuses on Human herpesvirus 4 and 6, two ubiquitous viruses with a long list of putative disease associations, ranging from malignancies such as lymphomas and carcinomas to multiple sclerosis. To date, the relatively limited genetic data on these organisms hinders the understanding of their variability and their genetic structure at a population level.

We here explore wet lab techniques for the production of genetic data in a cost-effective manner in order to reach the order of magnitude that is required to unravel the genetics these viruses. After successfully identifying individuals affected by integrated chromosomally inherited human herpesvirus 6 from public datasets of sequencing data, the sequences of the infecting virus were produced by target enrichment means from the source biological sample. The testing of an in-house target enrichment protocol followed, aiming to target latently infecting virus in human saliva. While the protocol still needs optimization and the aid of alternative techniques to be suitable for cost-effective, large-scale

studies, the results were very satisfactory (up to >800-fold enrichment). In parallel, long range PCRs were used to produce human herpesvirus 4 latency genes sequences from one large human healthy saliva panel including populations previously unexplored in terms of human herpesvirus 4 isolates.

Thanks to the combination of wet lab techniques and data analysis, the presence of genetic patterns in the two studied viruses is emerging, with human herpesvirus 6 presenting differences in diversity between its two species, as well as signs of geographical patterns possibly in part hidden by recombination events. Different bioinformatics approaches showed instead a stronger geographical stratification in human herpesvirus 4, with regional-driven clades.

This information would allow us for a correct study design when addressing the relationship between virus and disease, taking into account the natural variation of the virus, as well as help to pinpoint genetic features that might be determinant for disease triggering or development. The strong geographical patterns presented by the diseases associated

to these viruses strengthen the notion of the importance of this investigation and opens an avenue of research focused on disclosing the putative relationship between viruses strain variation and the risk for these virus–associated diseases.

Preface

We have walked a long road since the first discovery of those small biological and pathological agents that will become a staple in biological and medical studies: viruses. Nevertheless, the biology and the specific mechanisms that makes of certain viruses, pathological entities are yet mostly unknown. This thesis follows the traditional method of exploration by trial-and-error that can often place the bases for more detailed study in fields where much is yet to be understood. This exploration tackles the issue of understanding variability in complex organisms such as herpesviruses. These are viruses that have followed humans in their evolution since their separation from lineages leading to the other great apes, and that may have deeply influenced their biology and past and current pathologies.

We here take into consideration two herpesviruses in terms of populations, aiming to describe their variability as a proxy to understand their structure and stratification throughout the globe. We describe several wet lab techniques and

bioinformatics analyses designed to that effect and which include methodologies taken from DNA-capturing methods coupled to next generation sequencing to populations genetics, highlighting the best direction to continue studying these organisms and what could be expected from the results.

This work is embedded into the broader goal of building a path towards the understanding of viruses and associated diseases, an objective that will be obviously not covered in these studies, but that represent the direction to which the whole study is meant to push.

Index

2. INTRODUCTION

1.1. The history of virology: a huge tiny discovery

"...on opening the incubator I experienced one of those rare moments of intense emotion which reward the research worker for all his pains: at first glance I saw that the broth culture, which the night before had been very turbid was perfectly clear: all the bacteria had vanished... as for my agar spread it was devoid of all growth and what caused my emotion was that in a flash I understood: what causes my spots was in fact an invisible microbe, a filterable virus, but a virus parasitic on bacteria. Another thought came to me also, if this is true, the same thing will have probably occurred in the sick man. In his intestine, as in my test-tube, the dysentery bacilli will have dissolved away under the action of their parasite. He should now be cured."

-Félix d'Herelle, 1917

The origin of viruses is still unknown today, with theories spanning from the evolution of genetic elements that achieved the ability to move between cells, to the coevolution of acellular elements with their current cellular hosts. Wherever lies the truth, viruses are very old organisms that have in time shaped the history of the tree of life as we know it today.

While the discovery of viruses has been a trial and error process that started in the late 1800s, the first recorded appearance goes way back to a 4000-thousand years old Egyptian stele, where a poliovirus victim was inscribed. From there, records have been found in the rash on the mummified body of Pharaoh Ramses V, and in plague records from ancient Greece, Rome, and China, but they were usually attributed to malevolent or angry gods. Hundreds of years later, a different explanation was given. In 1886 Adolf E. Mayer published his findings on the "mosaic disease" of tobacco, a disease capable of passing between plants through filtered liquid extracts. He mistakenly attributed the disease to bacteria, and it was only after 6 years that the

Russian scientist Dmitri Ivanovski countered the theory. Ivanovski took advantage of the recently discovered "Chamberland-Pasteur filter", a ceramic filter with pores so small that it could be used to remove all bacteria or other cells known at the time from a liquid suspension. The interpretation of Ivanovski was that the agent must be a soluble toxin, and the conceptual leap of Martin Beijerinck was necessary to understand that the truth lay in a living organism. True enough, Beijerinck was convinced that the agent was a living organism of liquid nature, but the different view he gave to the study of viruses allowed the successive breakthroughs and the foundation of virology. Nevertheless, the true discovery of viruses is for many to be attributed to Friedrich Loeffler and Paul Frosch around 1898, when they performed similar experiments on calves (discovering the first animal virus in the process), but concluded that the infectious agent was not a liquid one, but a tiny particle instead. They even went further, heating infected vesicles extracts enough to inactivate its infectivity, and using them to vaccinate healthy cows. This is the first use of an inactivated virus as a prophylactic vaccine.

Subsequently, two discoveries started to put viruses in the limelight. The first was the one by Oluf Bang and Vilhelm Ellerman in 1908, when they associated virus with leukaemia by using a cell-free filtrate from diseased chickens to transmit the disease to healthy ones. The second was the discovery of the first virus associated with cancer by Peyton Rous in 1911, who demonstrated that the same disease transmission described by Bang and Ellerman was possible using extracts from solid tumour grafts. In his honour the virus was named Rous sarcoma virus. The events that made viruses become of global interest occurred between 1918 and 1922, when the worst plague of the 20th century, the Spanish Flu, lead to the death of 50 million people around the world.

It was in the 1930s that the invention of electron microscopy revolutionized virology by giving it a means to study virus structures. The sample preparation was still tremendously expensive due to the use of heavy metals such as gold, and only in 1959 the publication of "A negative staining method for high resolution electron microscopy of viruses" by Sydney Brenner and Robert Horne allowed the gathering, in just a

few years, of an enormous amount of information on viral structures. The technique was simple: a liquid virus sample would be deposited on carbon-coated metal grids and stained with moderately cheap heavy-metal salts, avoiding the use of expensive metals.

Virology became the science it is today when RNA was shown to be not only a conveyor of information between DNA and protein, but a replicable genetic material itself. This was discovered by Sol Spiegelman in 1965 for RNA, followed by the demonstration that the same direct replication could be used by a single strand DNA virus by Mehran Goulian in 1967.

Since then, many discoveries have followed, leading to the understanding of virus life cycles and especially of the interaction with the host, the mechanisms of cell inclusion, the immune response modulation and avoidance, and the genetic structure. While the biology of viruses was unfolded, the mechanisms of another important aspect of virology were emerging: virus evolution.

1.1.1 Main virus evolutionary paths, and influence in human evolution

Viruses are a very diverse group of organisms, showing tremendous polymorphism in almost all aspect of their biology, and evolution is not an exception. Nevertheless, this variability is restricted by the bond that viruses have with their host, a bond that can change deeply during viruses evolutionary history.

The more straight-forward process of evolution for a virus is to maintain the same host and to adapt through minor genome adjustments when new species branches out. This evolutionary path is known as co-divergence. Instances of this phenomenon can be found in our lineage, where we find different virus presenting very similar version for Humans and Chimpanzees (e.g. hepatitis B virus, or human herpesvirus 7(Hu, Margolis, Purcell, Ebert, & Robertson, 2000; Lavergne et al., 2014)). These viruses mutated from versions infecting these two species shared common ancestor more than four million years ago. This might seem as a long virus-host

relationship, but it is just an example of "recent" event. The *Herpesviridae* family itself for instance, has been co-diverging with vertebrates for 400 million years(Mcgeoch & Gatherer, 2005), and have separated during the way in avian, reptilian and mammals herpesviruses.

Co-divergence is not the only evolutionary path adopted by viruses, and as recently discovered not the most common. Viruses can jump from a species to an unrelated one, in a process called cross-species transmission. While the odds of a virus, an organism often very specifically adapted to a host, to manage to generate and maintain an infection and replicate in a completely different environment (i.e. host) are very scarce, this transmission mechanisms is common in viruses. A recent study by Geoghegan *et al*.(Geoghegan, 2017), 19 RNA and DNA virus families are analysed for co-divergence and cross-species transmission frequency. The result, well-resumed in Figure 1, shows how corss-species transmission is pervasive in most families, and while more frequent between more closely related species, it can even brake the vertebrate-invertebrate boundary.

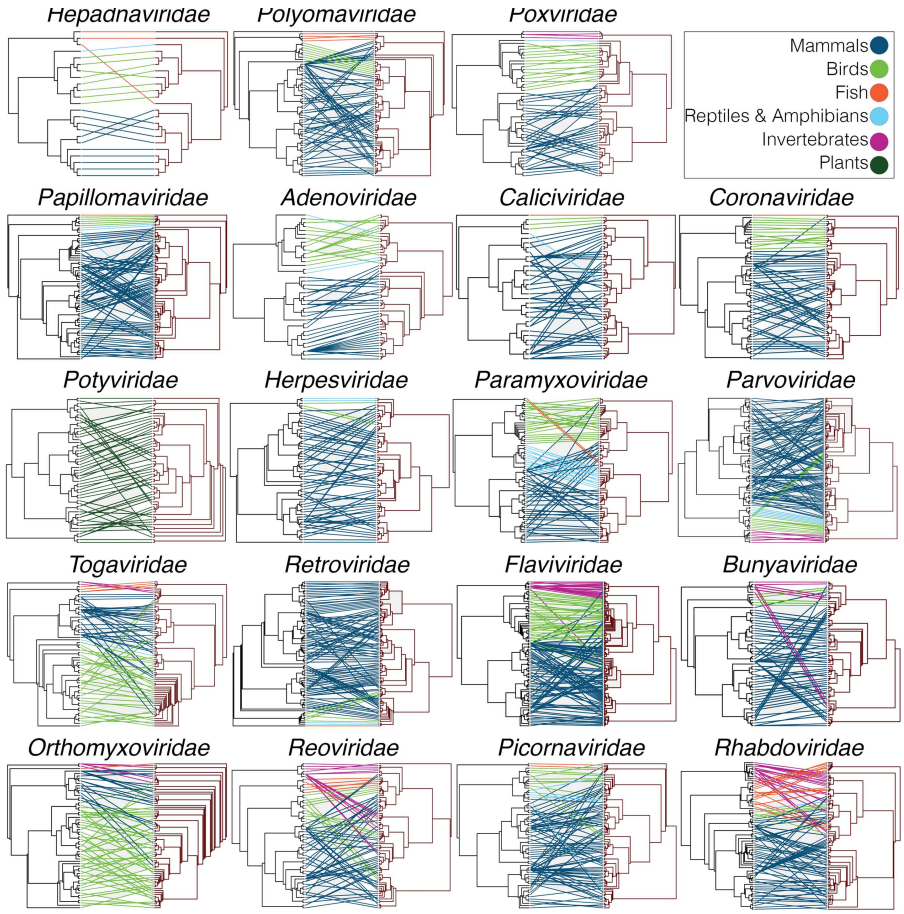


Figure 1. **Tanglegrams of rooted phylogenetic trees for each virus family.** Host trees were rooted first following their known phylogenetic history, with virus trees then rooted based on the host tree. The 'untangle' function was used to maximize the congruence between the host and virus phylogenies. Lines that connect the host (left) with its virus (right) are colored according to the host type (dark blue: mammals; light green: birds; light blue: reptiles and amphibians; red: fish; pink: invertebrates; dark green: plants). Figure and legend from Geoghegan *et al*.(Geoghegan, 2017)

RNA viruses appears to undergo cross-species transmission more frequently, possibly because of the very high rate of mutation and replication(Sanjuan, Nebot, Chirico, Mansky, & Belshaw, 2010). Also, the duration of infection is usually short in RNA viruses, limiting the development of co-divergence with the host. Cross-species transmission also has deep implications in virus-driven diseases in humans. The infection from virus for which there has been no time to adapt solid specific defence mechanisms can result in harsh pathological outcomes, such as the notorious SARS from a virus coming from bats, Dengue or HIV-1 from Old World primates or Chimpanzees, or Measles from cattle.

Co-divergence and cross-species transmission are not exclusive, viruses often adopt both paths at different moments of their history. Hominids have had a deep evolutionary relationships with human papillomavirus 16, with co-divergence shaping in large part the current virus populations and their geographical distribution. Still, cross-species jumps have been as fundamental in the distribution

and diversity of this viruses in Hominids(Pimenoff, Oliveira, & Bravo, 2016).

Virus evolution is thus not linear, and extremely complex in instances, and can influence its host evolution in different direct and indirect ways. The footprints of host-virus relationship in human evolution are in fact pervasive. We know today that at least 8% of our genome is made of active retroviruses or their remains(Horie et al., 2010), and a conservative estimate of the amino acid changes in the conserved part of the human proteome caused by adaptation to viruses goes above 30%. The most interesting aspect of this wide-spread adaptation, explored in Enard *et al*.'s work(Enard, Cai, Gwennap, & Petrov, 2016), is its target. It would be expected that the adaptation to virus would mainly drive adaptation in proteins with a specialized role in antiviral defence. While this has been proven for many antiviral defence proteins, for which exceptionally high rate of adaptation has been measured(Cagliani et al., 2012; Fumagalli et al., 2010; Kerns, Emerman, & Malik, 2008; Liu,

Wang, Liao, & Kuang, 2005; Sawyer, Emerman, & Malik, 2004, 2007; Sawyer, Wu, Emerman, & Malik, 2005; Sironi et al., 2012; Vasseur et al., 2011), most of the adapted protein have very diverse functions, mostly key ones in basic cellular functions. Furthermore, while the expectation would be to find host-virus co-evolution effects on less conserved genes, its action is more wide-spread. An useful and well-known example exemplifying this concept is the transferrin receptor TfR1(Demogines, Abraham, Choe, Farzan, & Sawyer, 2013). TfR1 is a cell-surface receptor that regulates a normal cell function, the intake of iron, but that also facilitates the entry of diverse viruses in mammals. The change of this receptor has been studied in rodents, where it is remarkably conserved, with the exception of few residues that show instead accelerated evolution. Remarkably, all of these residues correspond to regions of the protein that bind to viruses, and their mutation decrease virus entry without affecting the iron-intake capability of the receptor. This example, together with many others, reveals how virus-host co-evolution can unfold affecting even highly conserved and essential housekeeping genes.

It then becomes evident that understanding virus evolution is fundamental not only to counteract virus-related pathologies, but to understand human evolution itself.

1.2. Hidden within our cells: persistent virus infections and their implications

Viruses depend on cells for reproduction and metabolic processes. By themselves, viruses do not encode all the enzymes necessary for viral replication, but within a host cell, a virus can commandeer the cellular machinery to produce more viral particles. This characterizing feature defines viruses, the life cycle of which can be divided in two stages: a first stage of active and massive virion reproduction, and a second stage of minimum activity and avoidance of the host immune system. In this section, we will explore the two

stages with special regard to the second, which will be the focus of this thesis, as a means of persistent, life-long, infection.

1.2.1 Primary infection

In this stage, the virus is found in the form of virus particles, a protected structure that enables the transportation of the genetic material with the aim of spreading among cells and organisms. This process includes 5 main steps, achieved through a variety of means in the different virus families. It starts with the contact of the virus particle with the host cell (adsorption), followed by with its inclusion within the cell (penetration), the modification and induction of the cellular components to replicate the viral genetic material (viral genome replication), the reconstruction of the virus particle (maturation) and finally the exit from the cell (release).

Adsorption is one of the phases in which we find more diversity among viruses. Viruses tend to have exclusive or preferred cellular types for infection, and it is mainly the adsorption strategy that defines the virus tropism. Here, the

virus particle comes in contact with a selected host cell by recognizing receptors, often protein structures, on the cytoplasmic membrane. The choice of receptor is species-specific, and plays the main role in the virus tropism (*i.e.* the cell types that the virus can infect). The penetration phase consists in the viral components mediating the entrance of the viral genome in the cell. In viruses with animal hosts the penetration happens through endocytosis and, for enveloped viruses, through the fusion of the viral envelope with the host cell membrane. Naked virus can induce endocytosis too, or enter the cell by a poorly-understood mechanism of capsid rearrangement that ends with the viral nucleic acids released into the cytoplasm. Penetration is an intermediate phase between the contact and the beginning of the infection of the cell, but it is crucial for two reasons: the entry in the cell hides the virus from the extracellular immune system elements, and it is the moment in which the first viral genes begin to block viral-recognition mechanisms and start modifying the cellular replication machineries for the stages to follow. The viral genome replication phase exploits a variety of mechanisms depending on the type of virus, but mainly, even if in different

orders, its steps are the release of the viral genetic material and its replication, and the formation of the proteic parts needed for virion nucleocapsid assembly. This occurs in a crystallization-like process during the maturation phase. The nucleocapsid is a protein-based structure that will coat the virus genome, over which a large series of proteins will attach forming the tegument. The nucleocapsid shows remarkable geometrical structure that depends on its composition, and is a mean of classification of viruses. For DNA viruses, this stage implies the movement of the virus to the nucleus in order to access the cell replication components, while it usually takes place in the cytoplasm for RNA viruses. Very few large viruses, such as Pox viruses, produce DNA replication components themselves, and do not need to be in the nucleus in order to perform the replication. The entrance in the nucleus is usually mediated by the capsid, of which the interaction with a nuclear membrane pore induces a conformational change in the capsid itself and the release of the nucleic acids into the nucleus. As usual there are exceptions to these rules. For instance very small viruses can

enter the nucleus through a pore still covered by the capsid and release their nucleic acids only when inside.

The exit from the cell occurs in very different ways in enveloped and non-enveloped viruses. The envelope is a membrane that covers the virus particle for protection, is stabilized from within by skeleton proteins, and shows a variety of projections on the surface depending on the kind of virus. During maturation, a large number of non-enveloped virus accumulates in the nucleus or the cytoplasm, at which point the release phase begins. The textbook example of viral release is the one usually adopted by bacteriophages, a method shared by naked animal host viruses. In this strategy, the release is achieved through the sudden disruption of the cellular external membrane, killing the cell in the process, and liberating the large number of accumulated virions, even though even in this group it is not a rule. Enveloped animal host viruses are instead released individually during a variable interval of time. In this case, the release of the virus is part of the formation of the virion itself, which involves the direct budding-off from the cell membrane, or the exocytosis if

the envelope was achieved previously by budding from the Golgi apparatus or the endoplasmic reticulum. The released virus thus acquires a membrane that will be spiked with glycoproteins, completing the virion maturation. Newly formed viruses are then capable of crossing the extracellular space and to infect other cells.

1.2.2 Persistent infection

Viral infection can become persistent when the virus is not cleared from the host after primary infection, hiding instead in specific cells or kept at bay by the immune system. Persistent infection may involve stages of both silent and productive infection without rapidly killing or even producing excessive damage of the host cells. Even though the strategies of persistent infection are very diverse, they can be classified in three main groups: slow, chronic and latent infection. The slow infection is the only one that may lack an acute period of viral reproduction, and consists of a long incubation period and subsequent reproductive cycle progression, often concomitant with progression of the associated disease. In a

chronic infection, the virus presence is detectable but does not present the high levels of virus copies present in an acute infection, and can induce recurrent diseases. In latent infections, the activity of the virus ceases, and can reappear in episodes of acute infection. Different types of persistent infections can be employed by the same virus. For instance Epstein-Barr virus (EBV) undergoes latent infection in lymphocytes, but maintains a chronic infection in the pharyngeal epithelial cells in order to ensure the constant presence of virions in saliva, the main vector of transmission of this virus (Hadinoto, Shapiro, Sun, & Thorley-Lawson, 2009).

For a latent infection to persist, the viral antigen-driven immune response must limit the replication of the virus to a level that would limit tissue damage. Viruses that establish persistent infection must carefully avoid mechanisms that overwhelm immunity since they need a living host for their own survival. The balance between the viral strategy for maintaining persistent infection, the rashness of the immune response and the immunopathology that the infection itself

causes is a very delicate one, carefully orchestrated by the infecting virus. The mechanisms for maintaining this equilibrium are poorly understood, even though some virus strategies have been described in a more detailed manner (examples of different mechanisms can be found in the Herpesviridae family, and will be described in the following section). The main active role of the virus is the avoidance or modulation of the immune system, which is involves a variety of different strategies. Some viruses encode specific genes that target infected cells or the immune system such as human cytomegalovirus. This virus encodes for proteins able to degrade HLA class I molecules and interfere with the class II ones, actively reducing the capability of the cell to recognize foreign substances(Halenius et al., 2011). Another example is Epstein-Barr virus, which achieves the a similar effect by downregulating HLA class I expression(Griffin et al., 2013). Some viruses persist in certain cell types where they are hidden from the immune system, such as the infamous human immunodeficiency virus (HIV) or human herpesvirus 6. Some viruses limit their replication, thus limiting the antigen available to alert the immune system. Lastly, some viruses

are error prone in their replication, increasing the probability of generation of escape mutants, often found in small RNA virus. Most virus combine different techniques covering some or all of these avoidance mechanisms.

These viral mechanisms of immune evasion are balanced by mechanisms of regulation within the immune system. Evasion of the immune system is required to avoid eradication together with the limitation of the immunopathology that could affect the tissue and, as a consequence, the host. In particular, CD8+ T cells effector functions can cause a high level of tissue damage by killing infected cells and releasing inflammatory cytokines. Unsurprisingly, cytotoxicity and cytokine release are strongly decreased in these cell types during the majority of chronic infections, causing cell exhaustion, or the sustained expression of inhibitory receptors and the presence of a transcriptional state distinct from that of functional effector or memory T cells(Barber et al., 2006; Ejrnaes et al., 2006; Zaiac et al., 1998).

When viruses are unable to cause constant immune cell exhaustion they can undergo a unique transcriptional and

translational state in which the productive replication cycle, and thus the expression of most or all antigens, is silent but can reinitiate. This state is known as latency. While different forms of latency involve time-specific expression of sets of genes, and hence a dynamic pattern of exposed antigens, some are defined by a transcriptionally and antigenically quiescent state. This can be achieved in two different forms of the viral genome: proviruses and episomes. In the first case, the virus inserts its genome within the genome of a cell of the host. This process implies the complete fusion of the virus genome with the host one, with the consequent replication of the viral genome together with the host cell genome. Examples of this mechanism of latency can be found in HHV-6 (Morissette & Flamand, 2010a), or in the more notorious HIV, in which the expression of antigens is ceased, making it completely invisible to the immune system (Rasaiyaah et al., 2013). Alternatively, latency can be dictated by viral episomes, *i.e*. circular viral genomes that do not integrate into the host cell genome. Episome-producing viruses encode a structurally-related, sequence-specific DNA-binding protein that tethers the episome to the host

metaphase chromosome during mitosis. This allows the episome to be replicated by the host cell machinery without antigen expression. Examples of viruses adopting this form of latency are Kapopsi's sarcoma herpesvirus(Ballestas & Kaye, 2011; Uppal, Banarjee, Sun, Verma, & Robertson, 2014), or Epstein-Barr virus, as we will see in detail in the following section.

Persistent infection is restricted to a small niche of the human population, but in many cases, involves the majority of human beings, as shown in Table 1. The lack of known severe medical conditions directly caused by virus establishing persistent infection compared to more virulent non-persistent virus shifted in the past the study effort away from the former. Nevertheless, in the last two decades the increasing understanding of these viruses led to the association of chronic or persistent infections to many diseases (Table 1). The attention that emerged on these viruses because of the pathological consequence of persistent infections raises many evolutionary questions on the implication of these conditions.

			Disease during Chronic Infection		
Virus, Primary Nucleic Acid, Estimated Percent of Humans Infected	Major Site of Persistence (Organ or Cell)	Acute Infection Examples	Within Normal Hosts	Within Immunocompromised Hosts	References
Endogenous retroviruses (ERV), DNA, 100%	All	Not applicable	Unknown	Unknown	Seifarth et al., 2005; Virgin, 2007b
Anellovirus/Circovirus, DNA, 90%–100%	Many tissues	Unknown	Unknown	Unknown	Davidson and Shulman, 2008; Ninomiya et al., 2008; Hino and Miyata, 2007
Human herpesvirus 6 (HHV-6), DNA, >90%	Lymphocytes?	Roseola	Unknown	Meningoencephalitis, secondary infections, immunomodulatory?	Straus, 2000; Yamanishi et al., 2007
Human herpesvirus 7 (HHV-7), DNA, >90%	Lymphocytes?	Roseola	Unknown	Unknown	Straus, 2000; Yamanishi et al., 2007
Varicella zoster virus (VZV), DNA, >90%	Sensory ganglia neurons and/or satellite cells, lymphocytes	Chicken pox	Herpes zoster	Disseminated disease, hepatitis, pneumonitis	Zerboni and Arvin, 2008; Straus, 2000
Cytomegalovirus (CMV), DNA, 80%–90%	Myelomonocytic cells	Mononucleosis	Rare	Disseminated disease, vasculitis, pneumonitis, retinitis, hepatitis, gastroenteritis, meningoencephalitis	Mocarski et al., 2007
Epstein-Barr virus (EBV), DNA, 80%–90%	Pharyngeal epithelial cells, B cells	Mononucleosis	Burkitt's lymphoma, nasopharyngeal carcinoma, non-Hodgkin's lymphoma	CNS lymphomas, oral hairy leukoplakia, lymphoproliferative disease	Rickinson and Kieff, 2007; Straus, 2000; Kieff and Rickinson, 2007
Polyomavirus BK, DNA, 72%–98%	Kidney	Unknown	Unknown	Hemorrhagic cystitis (post bone marrow transplantation), nephropathy (post kidney transplantation)	Zur, 2008
Polyomavirus JC, DNA, 72%–98%	Kidney, CNS	Unknown	Unknown	Progressive multifocal leukoencephalopathy	Zur, 2008
Adeno-associated virus (AAV), DNA, 60%–90%	Many tissues	Unknown	Unknown	Unknown	Gao et al., 2004; Berns and Parrish, 2008; Schnepp et al., 2005a, 2005b; Chen et al., 2005; Erles et al., 1999; Blacklow et al., 1968
Herpes simplex type 1 (HSV-1), DNA, 50%–70%	Sensory ganglia neurons	Pharyngitis, encephalitis, keratitis,	Cold sores, encephalitis, keratitis	Increased severity of same diseases, pneumonitis, hepatitis	Straus, 2000
Adenovirus, DNA, up to 80%	Adenoids, tonsils, lymphocytes	Upper respiratory infection, gastroenteritis	Unknown	Enteritis, hemorrhagic cystitis, pneumonitis, hepatitis, others	Garnett et al., 2002; Wold and Horwitz, 2008
Herpes simplex type 2 (HSV-2), DNA, 20%–50%	Sensory ganglia neurons	Genital herpes	Genital herpes, encephalitis	Increased severity of same diseases	Straus, 2000
Kaposi's sarcoma herpesvirus (KSHV) or human herpesvirus 8, DNA, 2%–60%	Endothelial cells, B cells	Unknown	Castleman's disease, Kaposi's sarcoma	Kaposi's sarcoma, primary effusion lymphoma	Ganem, 2006
Hepatitis B virus (HBV), DNA, 350 million, ~5%	Hepatocytes	Hepatitis	Cirrhosis, hepatocellular carcinoma	Same diseases	McGovern, 2007; Rehermann and Nascimbeni, 2005
GB virus C, RNA, 1%–4%	Lymphocytes	Unknown	Unknown	Unknown	Stapleton et al., 2004; Berzsenyi et al., 2005
Papilloma virus, DNA, <5%	Epithelial skin cells	Unknown	Papilloma, cervical and other mucosal carcinomas	Increased severity and incidence of same diseases	Leggatt and Frazer, 2007; Howley and Lowy, 2007

Table 1. Examples of viruses known to chronically infect humans. Table adapted from Virgin *et al*. 2009(Virgin, Wherry, & Ahmed, 2009).

1.2.3 Evolutionary perspective on persistent infection

In an interesting paper, Lythgoe *et al.* discuss the effect of very rapid changes of viruses due to short generation time and high mutability. These characteristics are the foundations of the virus capacity to infect the host and maintain infection. Rapid change can however have a negative effect on the virus. With natural selection causing adaptation to the immediate environment, viruses can over-adapt to the host, which would be detrimental for the virus spreading to the rest of the population(Lythgoe, Gardner, Pybus, & Grove, 2017). As shown in HIV-1, HCV, and HBV(Kwei et al., 2013; Seki & Matano, 2012; Uebelhoer et al., 2008), the accumulation of mutations that are tailored to specific host genotypes, such as human leukocyte antigen types, can cause a substantial fitness cost in the absence of that specific immune response. This fitness loss is hypothesised to be the cause of the evolution of viral mechanisms that ultimately limit the changes in part of the viral population, such as integration within the host genome. Integration implies the co-reproduction with the cell using the cell machinery, causing the virus generation time to pair with the longer cell cycle, and to reduce strongly the mutability because of the robust DNA correction and

repair mechanisms of the cell. While part of the viral population stays silently integrated and evolves slowly, the other part can continue to adapt rapidly to the host.

The ability of viruses to evolve rapidly is one of the keys to their success, allowing them to evade host immune responses, evolve novel functions and explore new niches. This theory would explain how this is maintained without fitness loss, particularly in life-long infection, and the reason behind the evolution of mechanisms to slow it down.

Another evolutionary question that arises is how the high number of viruses integrated in our genome, or establishing persistent infection, could affect during ageing, or the process of ageing itself.

An example can be found in the *Herpesviridae*, human herpesvirus 5 (HHV-5), or human cytomegalovirus (HCMV). HCMV establishes chronic infection at an early age in 80-90% of the population(Pass, 2001) preferentially within undifferentiated myeloid lineages and monocytes(Hahn, Jores, & Mocarski, 1998). While mostly non-pathogenic in

healthy carriers, HCMV has been proven to cause primary or reactivation/recurrent infection in immunocompromised individuals. This results in severe diseases, including pneumonia, gastrointestinal disease and retinitis(Ljungman, 2002). More interestingly, epidemiologic studies in healthy elderly individuals showing that persistent HCMV infection induced reduced immune risk profile, a condition characterized by abnormal ratios of CD4/CD8 T cells(Olsson et al., 2000), poor proliferative response to polyclonal stimulation(WIkby et al., 2002), and generally a decreased immunocompetence towards other viral infections. Other studies showed in addition significant telomere shortening in T-cell populations bearing HCMV persistent infection(van de Berg et al., 2010), and the increase of relative and absolute counts of CD8 T-cells in the blood, with a decreased representation of the naive and the increased representation of the effector memory blood CD8 T-cells(Cicin-sain, Brien, Uhrlaub, Drabig, & Marandu, 2012). These impairments of the immune system lead to poor responses to virus infections, to vaccines, and poor life expectancy in the elderly(Olsson et al., 2000; G. . Wang et al., 2010). Taken together, these

evidences suggested the hypothesis that the magnitude of the effort of the memory T-cell pool to the control of HCMV infection could lead to the onset of immune senescence(Pawaelec et al., 2004).

In the light of this evidence, and keeping in mind that HCMV was only an example of the viruses possibly inducing immune senescence in their host, the non-pathological role of these virus must be redefined, taking into account the indirect consequences of the regulation of the immune system during persistent infection. Many of the effects of the persistent infection of most viruses is still poorly understood, raising the question on how large is the influence of viruses in ageing, or how the immune system has evolved to counteract this influence.

As a final note on the subject, the implication of the presence of viruses integrated permanently in our genomes is of great interest. In humans, at least 8% of the whole genome is estimated to be of viral origin (particularly, endogenous retroviruses)(Horie et al., 2010), but the presence of

integrated virus has been proven to be present and occur in almost all branches of the tree of life and during the whole of eukaryotic evolution(E.V, Senkevich, & Dolja, 2006). The contribution of integrated virus to the host evolution could then be significant, whether by introducing genetic variation and innovation, influencing the host gene expression, and domestication into new protein-coding genes with cellular functions.

Only the concerted effort of virology, genetics, evolutionary biology and paleovirology(Patel, Emerman, & Malik, 2011) will be able to answer these questions.

1.3. Epstein-Barr Virus and Human Herpesvirus 6

1.3.1 Brief introduction to the *Herpesviridae* family

Herpesviridae is a family of large DNA viruses with genomes ranging from 120-240 kbp. The members of the family share the capacity to establish life-long latent infections characterized by reactivation in immunocompromised patients. The name of the family derives from the Greek word *Herpein*, which translates to "to creep", and was given because of the skin lesions that appeared to creep or crawl on the skin of the patients. Nowadays, the name still applies to the family, fitting because of the chronic infection these viruses establish, keeping a slow but constant replication.

In virology, the main classifications are often based on the morphology of the virus, and herpesviruses are not an exception, sharing an enveloped core-capsid-tegument structure. The core consists of a single copy of a linear, double-stranded DNA molecule packaged at high density into the capsid. The capsid is an icosahedron, consisting of more than 160 capsomeres, each containing five to six copies of

the major capsid protein. The tegument, which surrounds the capsid, contains at least 30 viral protein species and is poorly defined structurally. In the tegument, structures positioned with symmetry corresponding to that of the capsid are detectable only in the region close to the capsid. The lipid envelope surrounds the exterior of the tegument, and is spiked with at least 10 viral membrane glycoproteins, in addition to some cellular proteins. The protein composition of the tegument and envelope varies widely across the family.

Herpesviruses have a life cycle consisting of an acute infection phase, followed by latent and chronic infections, and recurrent reactivations. Latency is achieved in the form of episomes and/or integration within the host genome. With the exception of Kapopsi's sarcoma herpesvirus and herpes simplex virus 2, all members of this family have very high prevalence in adult populations, even though geographical differences exist.

The members of the family infecting humans are divided in three subfamilies (Figure 2):

1- *Alphaherpesviruses*, herpes simplex virus 1 (HSV-1) and herpes simplex virus 2 (HSV-2), and varicella-zooster virus (VZV), characterized by a short replicative cycle, and the induction of cytopathology mainly in monolayer epithelial cell cultures. With their broad host range, these viruses are causative of known pathologies such as chickenpox, and oral and genital herpes.

2- *Betaherpesviruses*, HCMV, and HHV-6 and HSV-7, are characterized by a long replicative cycle and restricted host range. The only known pathology for which this subfamily is directly causative is *roseola infantum*.

3- *Gammaherpesviruses*, Epstein-Barr virus, or human herpesvirus 4 (HHV-4) and Kapopsi's sarcoma virus, or human herpesvirus 8 (HHV-8), show a very restricted host range, and are the members of the *Herpesviridae* proven to be carcinogenic.

Much can be told about this family of viruses, but for the sake of congruency, we will focus on the two species that are the study organisms in this thesis: HHV-4 (EBV) and HHV-6.

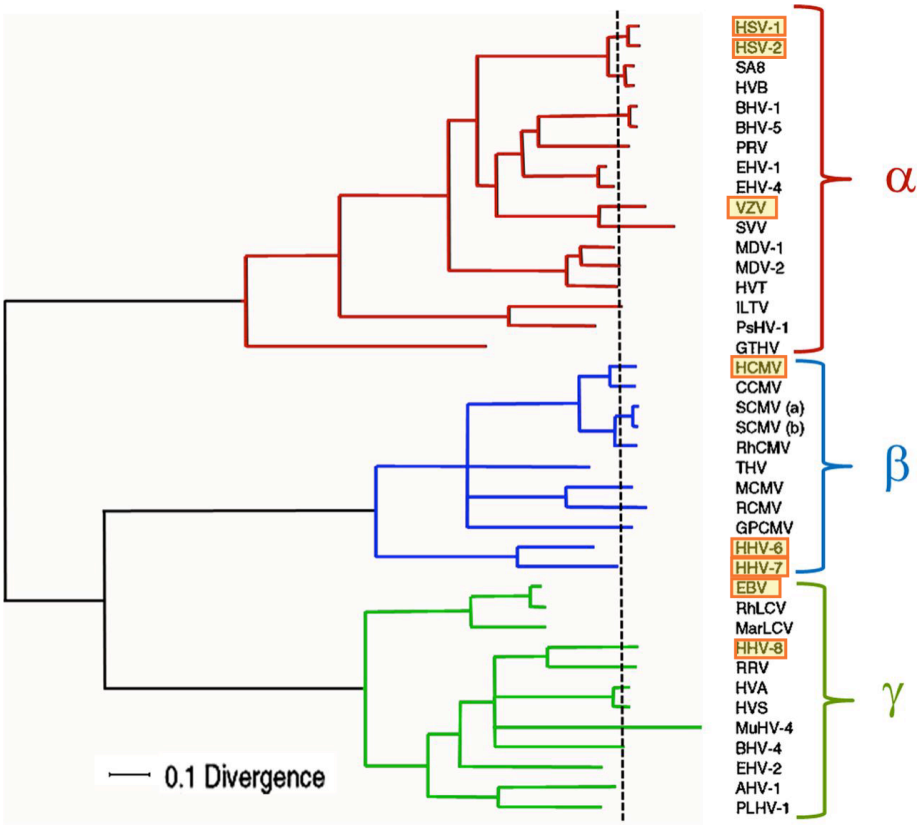


Figure 2. Composite phylogenetic tree of the Herpesviruses family. human-hosted herpesviruses are boxed in orange. Subfamilies are shown on the right, and subfamily members colored accordingly. Figure modified from McGeoch *et al*. 2006, adapted in Grose 2012 (Grose, 2012; Mcgeoch, Rixon, & Davison, 2006).

1.3.2 Human Herpesvirus 4 or Epstein-Barr Virus

HHV-4 is one of the most studied members of its family being the first human tumour virus discovered, and because of its

peculiar cell immortalization capability. It is an almost ubiquitous virus infecting the vast majority of adult human population(C. . Chang, Yu, Mbulaiteye, Hildesheim, & Bhatia, 2009) that causes infectious mononucleosis, and is associated to a diverse range of diseases.

Anthony Epstein was the physician who opened the door to the idea of a relationship between viruses and tumours with the discovery of a new herpesvirus in 1964: HHV-4. He developed the idea of a tumour can be caused by some factor related to the environmental conditions after listening to a talk given by Denis Burkitt on a lymphoma he was studying in Africa (what will become Burkitt's lymphoma). Epstein thought that the factor was related to the environment, and could be a biological one, and thus convinced Burkitt to send him samples of the African lymphoma to study.

For three years Epstein tried to understand what he hypothesized to be a biological factor linked to BL with no success. Only in 1963 a fortuitous event brought Epstein to his great discovery. When he decided together with his

collaborators Yvonne Barr and Bert Achong to try to culture BL cells to study their development, he had poor success. He was then surprised to see samples sent by Burkitt that had been delayed in their shipment, getting to his lab in a blurry transport medium. The cloudiness turned out to be actively replicating BL cells instead of the bacteria contamination he expected. Electron microscopy revealed that the cells were indeed infected by a biological agent: a virus. EBV had been discovered, and the first human lymphoma cell lines established. It was only in 1967 that the study proving that EBV was capable of immortalizing tumour and healthy lymphocytes was published. In later years EBV had been extensively studied, leading to discoveries such as the details of its life cycle and latency, its link with Hodgkin's lymphoma and nasopharyngeal carcinoma or the still debated role as a trigger for multiple sclerosis. Nevertheless, much is yet to be understood of the biology and pathological role of this virus.

EBV presents as two types(Zimber et al., 1986), named type-1 and type-2, with highly similar, colinear genomes. As the

rest of its family's genomes, EBV has a long linear double stranded DNA virome (ca. 172 kb), flanked by short terminal repeats (0.5 kb each), and divided in two unique regions by a series of internal repeats (Figure 3). EBV was the first large DNA virus to be completely sequenced(Baer et al., 1984). This first complete sequence was B95-98, the base of the present hybrid reference strain for the virus. B95-98 was found to be to be anomalous compared to the other EBV sequences for the lack of a 12 kb-long deletion. This was filled using another isolate (Raj strain)(Parker, Bankier, Satchwell, Barrell, & Farrell, 1990), completing the reference sequence. The complete EBV genome has 85-95 open reading frames, a number of which belong to the genes conserved in all *Herpesviridae* family.

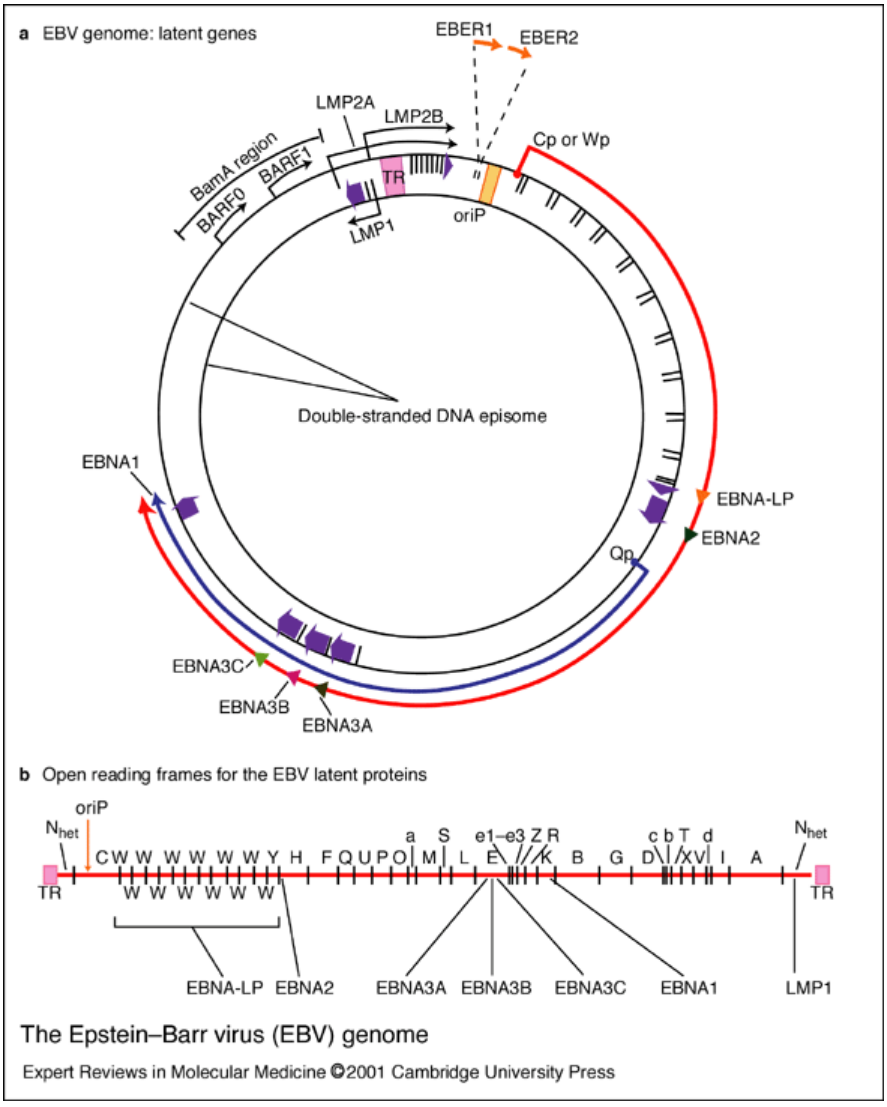


Figure 3. EBV genome and latency genes. (a) Diagram showing the location and transcription of the EBV latent genes on the double-stranded viral DNA episome. The replication origin of the plasmid (oriP) is shown in orange. The large solid blocks (in purple) represent coding exons for each of the latent proteins and the arrows indicate the direction in which they are transcribed. The orange arrows at the top represent the highly-transcribed latency RNAs EBER1 and

EBER2. The long outer arrowed red line represents EBV transcription during the main form of latency. (b) Diagram showing the location of the prototype B95-8 EBV genome ORFs for the EBV latent proteins using the classical *Bam*HI restriction endonuclease map nomenclature. Figure from Young & Rickinson 2004(Young & Rickinson, 2004).

EBV infects predominantly epithelial cells and B lymphocytes(Shannon-Lowe, Nauhierl, Baldwin, Rickinson, & Delecluse, 2006; Tugizov et al., 2007), where it establishes a lytic as well as a latent cycle. Latency is achieved in two forms: chronic infection of the pharyngeal epithelium and true latency in circular episomes bound to the host cell genome in lymphocytes. Being transmitted mainly through saliva between mother and child, or between children, the chronic infection of epithelial cells ensures constant virus shedding and presence of virions in this biological fluid(Hadinoto et al., 2009). The episomes ensure instead a life-long infection with the capacity of reactivation when the immune system is compromised.

EBV's life cycle is similar to the one of the rest of the herpesviruses, even though the productivity tends to be lower than its sister species, and the target are mainly B

lymphocytes. Primary EBV infection is nevertheless poorly understood since there is no cellular model where it has been studied. Lytic infection has only been studied by reactivating latent EBV or in transformed cell lines.

The entrance of EBV into the cell is achieved through two different routes. In epithelial cells entry is achieved by direct fusion with the membrane, while the virus needs to be endocytosed prior to membrane fusion in B cells(N. Miller & Hutt-fletcher, 1992; Nemerow & Cooper, 1984). In the first case, the CD21 or CD35 receptor interacts with the attachment proteins gp350/220 and BMRF2(Longnecker, Kieff, & Cohen, 2013). In the second case, five glycoproteins encoded by the EBV genome allow for cell entry, within which are the HHV-6-similar gH/gL-gB complex, and the receptor-binding gp42. gp42 is an important viral product for EBV because it defines through its interaction with gH/gL the specific tropism of the virus. This protein blocks the entry into epithelial cells, but together with gH and gL triggers the entry into B cells(Borza & Hutt-fletcher, 2002; J. Chen, Rowe, Jardetzky, & Longnecker, 2012). Human leukocyte antigens (HLA) class II molecules actively reduce the amount of gp42

in the budding virions in the host cell, but the epithelial cells lack this class of molecules. The virions produced in epithelial cells will thus contain high amounts of gH/gL/gp42 complex, which will make them highly infective for lymphocytes. In contrast, B cells virions contain little gp42 and have easier entry into epithelial cells(Borza & Hutt-fletcher, 2002; X. Wang, Kenyon, Li, Mullberg, & Hutt-fletcher, 1998).

When B cells lines are infected, EBV expresses a limited number of gene products that maintain virion production and protect the cell from apoptosis. The production of infective virions through lytic replication is recurrent in EBV hosts, and appears to occur in memory B cells circulating through the lymphoid tissue associated to the oropharyngeal mucosa(Faulkner, Krajewski, & Crawford, 2000). During lytic replication, large circular concatamers are produced as intermediates that are then cleaved and packed as new virions in the nucleus(Hammerschmidt & Sugden, 1988).

Latency is established by EBV in the form of episomes, circular forms of the virus genome that bind to the host DNA. Episomes are packaged into nucleosomes together with the

host DNA, and follow its behavior(Dyson & Farrell, 1985), undergoing a single replication for each cell division, and segregating uniformly into the new formed cells(Kirchmaier & Sugden, 1995). The attachment of the episome to the host DNA and its maintenance only requires the action of a viral product and its binding sequence: the Epstein-Barr Nuclear Antigen 1 (EBNA1), and OriP(J. Yates, Warren, Reisman, & Sugden, 1984; J. Yates, Warren, & Sugden, 1985). While there is no definitive proof, evidence strongly indicates that replication itself is carried out by the cell replication machinery(Adams, 1987; J. . Yates & Guan, 1991).

The switch between latent and lytic cycle is induced by a single protein that triggers a cascade effect. This protein, coded for by BZLF1, plays a role in viral replication, but most importantly it has key transactivator activity. Binding to the BZLF1-responsive elements present in the promoter regions of the immediate early genes, as well as many of the early genes, induces the formation of a viral initiation complex in a specific encoded sequence (oriLyt), and starts the lytic

cycle(Gao et al., 1998; Hammerschmidt & Sugden, 1988; Packham, Economou, Rooney, Rowe, & Farrell, 1990).

Latency is established in 1 to 50 B lymphocytes per million(Babcock, Decker, Freeman, & Thorley-lawson, 1999). It is regulated by only 11 proteins and a few non-coding RNAs, within which are a series of miRNAs. Notably, EBV is the only known virus to encode miRNAs(Pfeffer et al., 2004). Latency genes and functions are listed in Table 2.

Gene product	ID	Class	Function
Epstein-Barr nuclear antigen 1	EBNA1	Protein	-Latency-lytic switch -Essential for episomal maintenance and replication
Epstein-Barr nuclear antigen 2	EBNA2	Protein	-Necessary for B cells transformation
Epstein-Barr nuclear antigen 3	EBNA3	Protein	-3 products: EBNA3A, 3B, 3C -Essential for B cell transformation -EBNA 3B may function as tumor suppressor in B cell lymphomas
Epstein-Barr nuclear antigen Leader Protein	EBNA-LP	Protein	-Required for LCLs outgrowthme

Latent membrane protein 1	LMP1	Protein	-Oncongene -Essential for B cell transformation
Latent membrane protein 2	LMP2	Protein	-2 products: LMP-2A and -2B -Can rescue B cells from apoptosis
BamHI-H rightward reading frame 1	BHRF1	Protein	-Protects from apoptosis in Burkitt lymphoma
BamHI-A rightward reading frame 1	BARF1	Protein	-Expressed in nasopharingeal carcinoma and gastric cancer -Induces transformation of B cells in mice and humans
BamHI-A rightward transcript miRNAs	BART miRNAs	miRNAs	-44 mature transcipts -Maintain latency -Can drive tumour growth
Bam HI-H rightward transcript miRNAs	BHRF1 miRNAs	miRNAs	-3 transcripts -Important in B cells transformation
EBV-encoded RNAs	EBERs	RNAs	-Highly abundant -Activate innate immunity

Table 2. EBV latency genes products and functions.

Epstein-Barr virus is capable of immortalize B lymphoblastoid cells leading to the establishment of polyclonal, continuously growing cell lines(R. . Chang, Fillingame, Paglieroni, & Glassy, 1976; Henle, Diehl, Kohn, Zur Hausen, & Henle, 1967; G. Miller, Lisco, Kohn, & Stitt, 1971). The essence of its biological behaviour is that it initiates, establishes, and maintains persistent infection by subtly exploiting various aspects of normal B cell biology and has evolved to minimally

perturb the normal behaviour of the infected B cells. The mechanism mimics natural B cell activation, as represented in Figure 4. This mechanism, related to tumoural induction in case of disruption, has and is still largely used to immortalize laboratory cell lines.

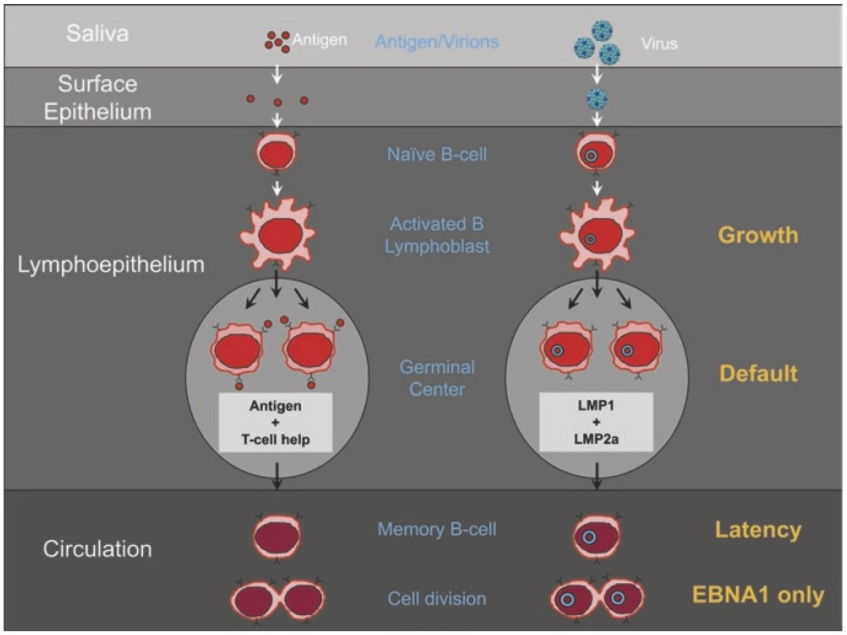


Figure 4. EBV biology mirrors B cell biology. To the left is schematized a typical mucosal humoral immune response. Antigen in saliva crosses the epithelial barrier of the tonsil to be sampled by naïve B cells in the underlying lymphoid tissue. When naïve B cells recognize cognate antigen, and receive signals from antigen-specific T helper (Th) cells, they can leave to become resting memory B cells that occasionally undergo division as part of memory B cell homeostasis. To the right is schematized how EBV uses the same pathways.

EBV is spread through saliva, crosses the epithelial barrier, and infects naïve B cells. These become B cell blasts until the viral latent proteins LMP1 and LMP2 have the capacity to provide antigens and Th survival signals that allow the latently infected B cells to induce EBV to exit from latency and become resting memory cells that also divide through homeostasis. To the right are listed in orange the transcription programs used at each stage. The blue circles represent the viral DNA which is a circular episome. Figure from: Epstein-Barr Virus Volume 1(Münz, 2015).

EBV has been related to different pathologies since its discovery in Burkitt lymphoma samples. The virus is an etiological factor for Infectious Mononucleosis, commonly known as the kiss disease, a syndrome described in 1920(Sprunt & Evans, 1920) where activated T lymphocytes respond to EBV infected cells causing fever, pharyngitis and adenopathy. The duration and age of acquisition of EBV infection is relevant since it is related to the development of an asymptomatic infection or a symptomatic one (with the appearance of infectious mononucleosis), and to the susceptibility to certain malignancies and autoimmune diseases. Relevant examples are the increased risk of Hodgkin's lymphoma among individuals who have

experienced infectious mononucleosis(Hjalgrim et al., 2000), or the higher risk of developing multiple sclerosis(Ascherio & Munger, 2010). Acquisition of EBV at an early age is also postulated to increase susceptibility to different pathologies, such as nasopharyngeal carcinoma(Melbye, Ebbesen, Levine, & Bennike, 1984) or Burkitt lymphoma(Piriou et al., 2012; Slyker et al., 2013). Age of primary infection is thus an important factor for the IM epidemiology, and important for EBV-related diseases.

The mechanism of infection and activation of B cells by EBV can cause malignancies under certain circumstances. Over 1% of all worldwide human cancer has been shown to have EBV as causative agent. EBV-related neoplasia of lymphoid origin can develop into Burkitt or Hodgkin's lymphomas, a causative relation that has strong environmental patterns. In underdeveloped countries EBV is detected in a very high proportion of Hodgkin's lymphoma cases, compared to less than half of those of developed countries(Flavell & Murray, 2000). A similar pattern is found in Burkitt lymphoma, were almost the totality of cases occurring in countries where malaria is common are etiologically linked to EBV(Brady,

MacArthur, & Farrell, 2007). Less documented are the EBV-related neoplasias of epithelial origin. Here we can find gastric carcinoma, with 10% of the total cases being EBV-positive(Iizasa, Nanbo, Nishikawa, Jinushi, & Yoshiyama, 2012), and nasopharingeal carcinoma, with the vast majority of cases occurring in undeveloped countries being EBV-positive, similar to Burkitt and Hodgkin's lymphomas(Raab-Traub, 2002). These latter cases are even more relevant in their relation with EBV since they are thought to be derived from a single EBV-infected epithelial cell(Pathmanathan, Prasad, Sadler, Flynn, & Raab-Traub, 1995; Raab-Traub & Flynn, 1986).

Lastly, EBV has been associated to autoimmune disorders, with multiple sclerosis the first and most relevant example. While there is no direct proof of the relation between the virus and multiple sclerosis, there are many observations indicating its existence. These vary from similar geographical patterns, to EBV-positivity, patients affected by infectious mononucleosis and multiple sclerosis later in life, epitope specificity, EBV-specific antibody titration, similarity of the demyelinization plaques to virus-induced ones and the

presence of EBV in the central nervous system(Alotaibi, Kennedy, Tellier, Stephens, & Banwell, 2004; Ascherio & Munger, 2010; Banwell et al., 2007; DeLorenze et al., 2006; Levin, Munger, O'Reilly, Falk, & Ascherio, 2010; Pakpoor et al., 2013; Pohl et al., 2006; Serafini et al., 2010). Infecting the majority of the adult human population, and being causatively related to different diseases, EBV becomes a very relevant health issue that needs to be understood in its totality. The strong pattern of variation between age, environment and diseases calls for a deep analysis of the genetics of this virus and its stratification.

Pathologies associated to EBV show strong geographical patterns, which could be driven by genetic variation of the host or local environment factors or co-infections, as well as the natural EBV variation itself. The substantial geographic variation in the virus sequence in normal infected populations leads to the hypothesis that endemic strains of EBV in some parts of the world might be inherently more able to contribute to cancers or other EBV-related pathologies. Distinct

polymorphisms between EBV strains originating from different populations is not unlikely due to the early age of contact with the virus, and the life-long infection. Furthermore, instances of EBV polymorphisms at different frequencies in different populations has been shown. Campos-Lima *et al*.(de Campos-Lima et al., 1993), for instance, described population-specific alleles of the EBNA4 gene. An 8 peptide residues region of this gene is recognized by the antigen HLA-A11 in a viral infection control process. HLA-A11 is infrequent in Caucasian and central African populations, while it is very frequent in New Guinea, and EBV consequentially presents high frequency of a non-synonymous mutation in one of the peptides of the HLA-A11 binding site. This is a clear example of how the stratification of EBV can be related to the genetic architecture of the host population, and an additional evidence of the influence of the virus-host co-evolution consequences on the viral genome.

Most of the described EBV strains came from cancer cell lines or have been selected by B cell transformation to make

an Lymphoblastoid Cell Lines (LCLs), limiting the analysis of natural variation. Santpere *et al.* showed in a limited data set the divergence between Asian and African sequences in reportedly healthy individuals(Santpere et al., 2014), which was confirmed by an analysis based on a larger part of the genome by Tsai et al and Kwok *et al.*(Kwok et al., 2014; Tsai et al., 2013). In all analyses the divergence between types remains determinant, even if mainly driven by the hypervariable EBNAs(Rowe et al., 1989; Sample et al., 1990; Sculley et al., 1989). The frequency of the two types has a pattern itself, with type 1 being found relatively homogeneously world-wide and type 2 showing instead high frequencies mainly restricted to Sub-Saharan regions. While presenting near-equivalent biological properties, the main known difference between types is in the capability of immortalizing B cells, where type 1 shows faster and more effective induction of the process(Lucchesi et al., 2008).

In 2015 Palser *et al.* described 71 new EBV sequences from different tumour types or LCL, and the first healthy individual saliva-derived strain(Palser et al., 2015), deeply improving

the limited number of whole genomes available, and allowing for high resolution analysis. A first result confirmed the representation of the B95-98 hybrid genome of EBV in the whole world, supporting its role as reference sequence. The data also highlighted the marked difference between type1 and type2, mainly relating it to the variation of EBNA2 and EBNA3 genes. Noticeably, intertypic recombination was found in 2 cases.

The samples used for Palser's study originated from different regions around the world, but again the main difference evident was the separation between Asian and African individuals (Figure 5).

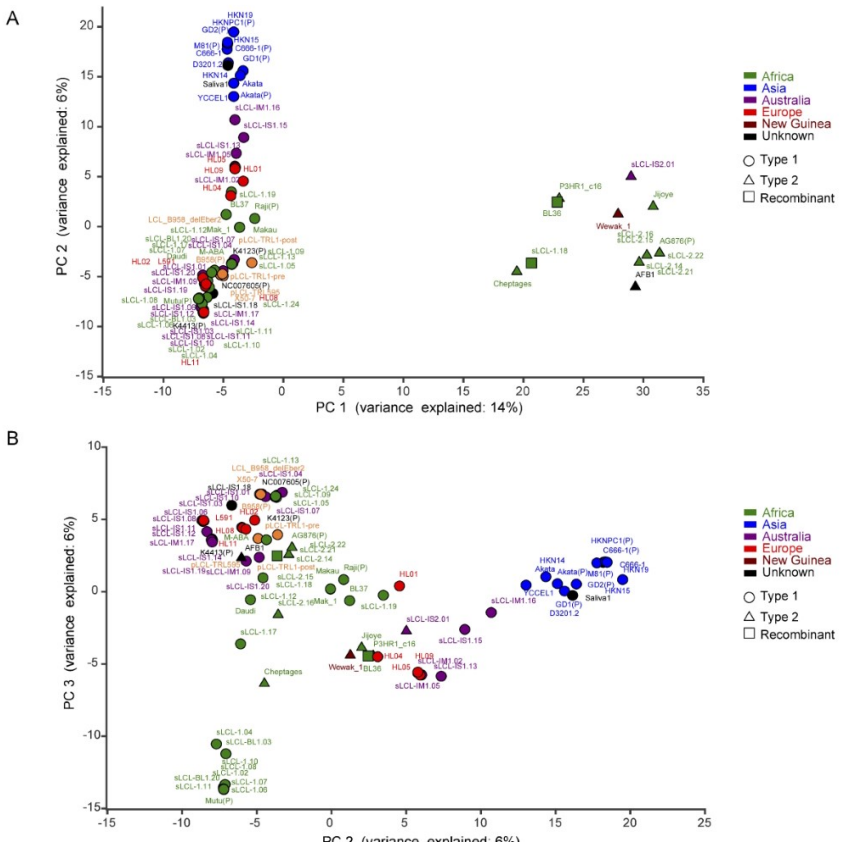


Figure 5. PCA analysis of 71 whole EBV genomes from different geographic origins. While small clusters can be seen between sequences with diverse origins, the main separation is between Asian and African individuals. From: Palser *et al.* 2015(Palser et al., 2015).

Lastly, the study by Palser *et al*. supports the higher variability of latent genes compared to lytic-related ones when analysing intra-strain variation.

Palser *et al*. data set has been expanded by the addition of Multiple sclerosis-derived sequence produced by Matteo *et al*.(Matteo et al., 2016) In this study a total of around 130 sequences is analysed for recombination patterns, and a strong stratification is found in EBV populations when the "pure" (*i.e*. non recombinant sequences) are used to built a phylogenetic tree (Figure 6).

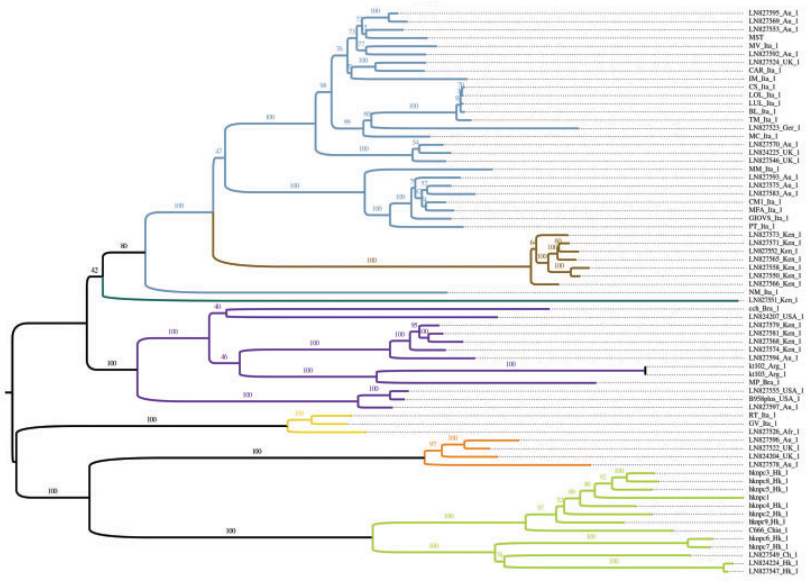


Figure 6. Phenetic tree of the “pure” genomes. Different groups are indicated by colors and the root position is arbitrary. “Pure” genomes are defined as those where Structure assigned a 90% probability of provenance from a single population. Figure and legend adapted from Matteo *et al*. (Matteo et al., 2016)

The geographic structure that appears in Matteo's et al. analysis is a good sign of EBV's stratification, but limiting the use of non-recombinant sequence would strengthen such stratification signal, magnifying the separation in the tree.

It is important to take in consideration that the geographical distribution of different EBV strains and EBV-related diseases is not only driven by genetics, but by other factors. Within these, the environmental factors and hygiene conditions stands out, and could help shaping the different prevalence and incidence of diseases such as multiple sclerosis or Burkitt's lymphoma.

In order to understand the natural variation to EBV, which could be an important factor to identify the mechanisms that generate the relation between the virus and its related diseases, more sequences are needed from more populations word-wide, as well as from healthy carriers. This takes higher importance when considering the strong prevalence and incidence patterns of the diseases associated

to the virus, which could reflect virus specific population variations.

The viral sequence found in healthy individuals would highlight the variation between natural EBV and the potentially selected EBV strains that induce diseases and are thus sequenced in pathology-related samples. The difficulty in sequencing EBV from healthy individuals is due to the low viral load that the virus presents in latency, but developing enriching or amplifying methods to achieve this goal would allow to be able to pick from an enormous number of hosts for sampling (approximately 90% of the world adult population).

1.3.3 Human herpesvirus 6

Human herpesvirus 6 (HHV-6) is classification composed by two different species, HHV-6A, and HHV-6B (D. Ablashi et al., 2014), that infects the majority of the adult human population (Ihira et al., 2002; Okuno et al., 1989; Saxinger et al., 1988). The two viruses had been traditionally considered as one

since their discovery in 1986 (Salahuddin et al., 1986) in samples from lymphotropic disorder patients. While sharing many aspects of their life cycle and structure, HHV-6A and -6B show specific differences in molecular, epidemiological and biological properties (D. V. Ablashi et al., 1991; Aubin et al., 1991; Schirmer, Wyatt, Yamanishi, Rodriguez, & Frenkel, 1991; Wyatt, Balachandran, & Frenkel, 1990). The difference became clear also genetically when two isolates were sequenced from AIDS patients and published(Dominguez et al., 1999; Gompels et al., 1995): U1102 (from Uganda), and Z-29 (from Zaire). These two viromes remain to date the reference sequence for HHV-6A and -6B, respectively.

The genome of HHV-6 consists of linear, double-stranded DNA of approximately 160 kbp. Two 7-8 kbp direct repeats (DRl, DRr) flank a unique region where the genes are densely compacted. Three additional repeats, named R1, R2 and R3, interrupt the unique region in the part where the immediate-early genes are encoded (Figure 7a).

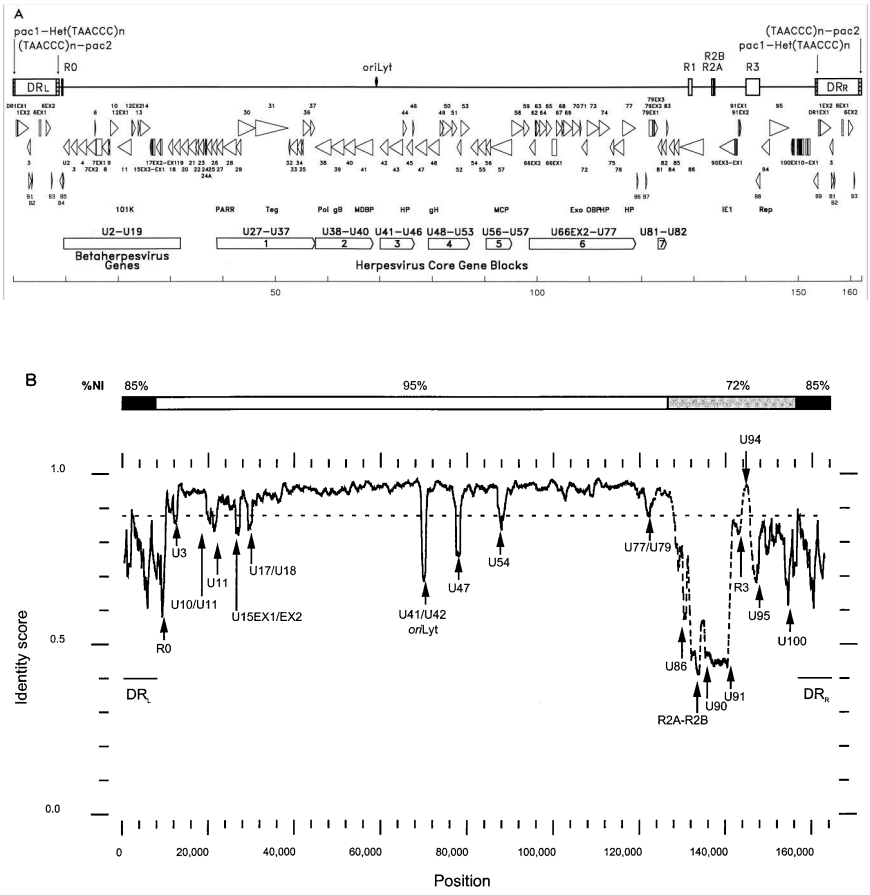


Figure 7. A) HHV-6B genome representation. The HHV-6A genome follows an almost identical organization, with the lack of the "B" genes found in HHV-6B mostly in the large genome-flanking direct repeats. B) Nucleotide sequence comparison between HHV-6A and HHV-6B. The horizontal dashed line represents the mean identity across the whole alignment. Low-identity ORFs are identified by their name (UX, where X is a progressive identifying number). Figure from Dominguez *et al*. 1999 (Dominguez et al., 1999).

The direct repeats contain the cleavage motifs pac-1 and pac-2(Thomson, Dewhurst, & Gray, 1994), and a variable number of a perfect or imperfect (GGGTTA) repeat that is identical to the human telomeric repeat sequence (TRS) (Meyne, Ratliff, & Moyzis, 1989). The two HHV-6 species (a.k.a. variants) contain 110 ORFs with colinear positions, plus 9 variant-specific ORFs for each variant (Dominguez et al., 1999; Gompels et al., 1995).

The two variant genomes have an overall 90% nucleotide identity, which varies from 98-99% in the central conserved regions, to 70% in variable regions such as the IE-A, with peaks that go lower than 50% (Figure 7b).

The HHV-6 life cycle follows its family's general *modus operandi*, where the virus generates a primary infection and undergoes latency as chronic and persistent infection with occasional reactivations. The cycles start with the transmission of the virus, which is mainly achieved through saliva. This discovery was driven by the frequent detection of HHV-6 in salivary gland tissue and saliva (Collot et al., 2002;

Di Luca et al., 1995; Zhao, Fan, Mu, Shen, & Cheng, 1997), from where the virus would spread from mother to child, or between children. It is of note that all isolates from saliva were found to belong to the B variant. Several studies shows that HHV-6A, as well as HHV-6B, could be transmitted sexually (Leach, Newton, McParlin, & Jenson, 1994; Zhou, Chang, Qian, Chandran, & Wood, 1994).

After transmission, HHV-6 establishes lytic infection preferentially in $CD4^+$ T lymphocytes *in vitro* (Lusso et al., 1988) and *in vivo* (Takahashi et al., 1989). While HHV-6 isolates were found capable of infecting peripheral or cord blood mononuclear cells, differences are present between the two viral variants. HHV-6A for instance is the only one capable of infecting $CD8^+$ T cells *in vitro* (Grivel et al., 2003). The tropism range *in vitro* is not only limited to T cells, but includes epithelial cells (Zhou et al., 1994), fibroblasts (Luka, Okano, & Thiele, 1990), immature hematopoietic cells (Furlini et al., 996), hepatocytes (Inagi et al., 1996), and astrocytes (T Yoshikawa, Asano, Akimoto, et al., 2002). Lastly, monocytes and macrophages can be infected by HHV-6, but there is a

controversy on the productive nature of such infection (Kakimoto, Hasegawa, Fujita, & Yasukawa, 2002; Smith et al., 2003). The *in vivo* tropism of HHV-6 is even broader, including brain tissue(Chan, Ng, Hui, & Cheng, 2001), liver tissue (Harma, Hockerstedt, & Launtenschlager, 2003), tonsillar tissue (Roush et al., 2001), salivary glands (Fox, Briggs, Ward, & Tedder, 1990), and endothelium (Caruso et al., 2002). Bone marrow progenitor cells ($CD34^+$) also have been found to be infected by latent HHV-6, which can be transmitted longitudinally to differentiated blood cells of different lineages (Luppi et al., 1999).

The broad tropism of HHV-6 is consistent with the ubiquity of its cellular receptor: CD46 (Santoro et al., 1999). In both HHV-6 variants the protein complex composed of the three viral glycoprotein gH, gL and gQ mediates the membrane fusion and entry in the cell (Mori, Yang, Akkapaiboom, Okuno, & Yamanishi, 2003). After fusion of the envelope with the cell membrane, the viral nucleocapsid is transported through the cytoplasm to the nucleus, passing through the nuclear pores. Here the virus promotes the sequential

expression of three classes of viral proteins that are named based on their time of appearance. The first are the immediate early proteins, which are synthetized within the first hours of infection, and orchestrate the expression of other genes preparing the next stage of the expression program. The early genes are then expressed, starting the replication process and managing the viral metabolism. Lastly, the late genes, which mostly encode components used for maturation of the new viral particles, are expressed.

As most of the other *Herpesviridae*, the replication of HHV-6 starts with the formation of a denatured gap in the origin of the lytic replication region (ori-lyt), from where the genome will be copied thanks to the combined action of viral and cellular molecules. The replicating genome is in circular form, generating long concatamers that are separated during the encapsidation process. The new capsids bud out of the nucleus acquiring a temporary membrane that will be covered by the tegument and a spiked envelope at the Golgi complex, where the viral glycoproteins accumulates. Finally, the viral particles are encapsulated in transport vesicles that are

released from the cell by exocytosis. HHV-6 takes approximately 72 hours to get from entry into the cell to the release of the new virions (Black et al., 1989).

Following the lytic cycle, HHV-6 establishes latency in order to persist in the host indefinitely. HHV-6 latency involves a chronic infection, characterized by low-level replication in salivary glands (Fox et al., 1990; Jarrett, Clark, Josephs, & Onions, 1990) and brain tissue (Chan et al., 2001; Donati et al., 2003), and integration into the host genome with no virion production. The integration, which takes places in very few cells of the body, is achieved via homologous recombination of the TTAGGG repeat encoded within both large direct repeats that flank the HHV-6 genome. These repeats mirror the same pattern present in the human telomeric and subtelomeric region, ending in the insertion of the viral genome at the junction of the two (Arbuckle et al., 2010).

HHV-6 infrequently integrates into the germ-line and can be transmitted vertically in a Mendelian fashion. The resulting condition is characterized by the presence of a copy of the virus in each cell of the body, in a form defined as inherited

chromosomally integrated HHV-6 (iciHHV-6), or congenital HHV-6. This condition affects approximately 0.2-1% of the adult population worldwide (Griffiths et al., 1999; C. B. Hall et al., 2008; Leong et al., 2007; Tanaka-Taya et al., 2004). This form of the virus appears to diverge from the non-integrated form (exogenous) and deep characterization of this divergence (Joshua Tweedy et al., 2016), and of the different pathologies that might be associated with the two forms, are under study.

During its life cycle, HHV-6 regulates and interacts with different parts of the immune system in order to avoid its eradication from the host. The virus has been shown to alter immune-mediated transmembrane signalling, altering deeply the functionality of cells belonging to the major T lymphocyte subpopulations (Lusso et al., 1991). Another broad-range interaction of this virus with cell surface molecules is the downregulation of CD46, a protein complex almost ubiquitous in human cells (Santoro et al., 1999) and that is a necessary element for HHV-6 to enter the cells. Other regulatory effects

involve different classes of chemokines (Hasegawa, Yasukawa, Sakai, & Fujita, 2001; Yoshida et al., 1997), molecules that have diverse roles, including the involvement in inflammatory processes, and the recruitment of cells of the immune system to the site of infection. HHV-6 itself also encodes chemokines with diverse potential effects. Within these there are the regulation of T-cell activation, and of monocyte- and macrophage-dependent inflammation(Milne et al., 2000). HLA molecules have also been reported to be downregulated by HHV-6, suggesting immune evasion strategies(Hirata, Kondo, & Yamanishi, 2001). Lastly, while inducing the production of interleukin beta 1 and tumor necrosis factor α, HHV-6 suppresses T-cell proliferation and function, as shown by the reduced interleukin 2 production and up to 85% reduction in their mitogen-driven proliferative response (Flamand et al., 1991; Flamand, Gosselin, Stefanescu, Ablashi, & Menezes, 1995).

This evidence suggests the potential of HHV-6 to interfere and modulate different major compartments of the immune

system, which could contribute to the spectrum of pathologies associated to this virus.

HHV-6 is widespread through the world, with a prevalence that averages 95% of the populations, but fortunately it is only etiologically linked to *roseola infantum* in its HHV-6B variant. *Roseola*, or *exanthema subitum*, is a disease that usually affects young children, and is characterized by high fever, diarrhea, and mild skin rash (Yamanishi K., Okuno T., Shiraki K., Takahashi M., Kondo T., Asano Y., 1988).

After establishing persistent infection, HHV-6 has the ability to reactivate from latency in immunocompromised patients, and recreate an active infection. Reactivation involves the transcriptional activation of IE genes triggered by unknown factors and the subsequent transactivation of cellular and/or viral factors. An active viral infection that is usually contained by the immune system in a healthy individual can have serious consequence in an immunocompromised patient. For instance, solid organ or bone marrow transplant patients have been shown to often present with encephalitis, pneumonitis,

and even bone marrow suppression or graft rejection due to HHV-6 reactivation (Braun, Dominguez, & Pellett, 1997)[-64].

The capability to reactivate when the immune system is weak or over-charged, make HHV-6 a virus difficult to associate to disease. The usual first clue to an association is the presence of higher titres of the virus in disease patients compared to healthy individuals. A major conundrum is to determine whether high HHV-6 viral loads are due to actual association, or just to the reactivation caused by the compromised immune system of the patient. Nevertheless, the relationship between different diseases and HHV-6 has been studied more in detail, even though a definitive proof of disease association in most of them is yet to be found. The more definite associations range from multiple sclerosis (Berti et al., 2002; Sola et al., 1993; Soldan et al., 1997), to meningitis (K. Yao et al., 2009; Tetsushi Yoshikawa et al., 2009), to Hodgkin's lymphoma (Maeda et al., 1993; Siddon, Lozovatsky, Mohamed, & Hudnall, 2012; Strenger, Aberle, Nacheva, & Urban, 2013) (for which it is worth mentioning

that the Bell *et al*. study did not find signs of association(A. J. Bell et al., 2014)).

The disease associations to the inherited chromosomally integrated form of HHV-6 are poorly known due to the relatively short time that it has been the centre of a larger scientific effort aimed to its understanding. Also, the time-consuming and complex techniques used to detect iciHHV-6 have only recently been replaced with an efficient ddPCR detection method, allowing for fast and cost-effective screenings (Sedlak et al., 2014). The consequence is a lack of whole iciHHV-6 genomes that hinders the genetic study of the virus and consequently of its disease associations, even though during the last few months different studies have shown the results on the production of new sequences. Lastly, also for the exogenous form of HHV-6 there are few described whole genomes (5 complete published genomes between the two variants), which are needed to define the possible difference between the viral forms, and permit correct study designs to understand the associated pathologies. Nevertheless, iciHHV-6 has been systematically

detected in a higher percentage of patients than controls for many of the associated pathologies, namely *angina pectoris*, thanks to a large-scale study by Gravel *et al*. in 2015 (Gravel et al., 2015).

The integration of HHV-6 takes place at the subtelomere-telomere junction, and it has been found to cause telomere shortening in somatic cells (Huang et al., 2014). The mechanism of the induction of such shortening is yet to be understood, but one possible explanation involves the alteration of repeat-encoding RNA expression. These transcripts are known to contribute to telomere shortening (C. Wang, Zhao, & Lu, 2015). The long-term consequence of a shortened telomere could cause premature senescence or cell apoptosis when numerous cell divisions are required.

Overall, the pathologies associated to HHV-6 and iciHHV-6 are mostly yet to be recognized, underscoring the need to understand the multiple confounding factors hampering correct study designs. Within these the most relevant are the differences between exogenous, integrated and inherited HHV-6, their variability and integration outcomes, and the

possible population stratification, raising the necessity of large numbers of complete viral genome sequences.

As for EBV, to understand the natural variation in healthy carriers could lead to the identification of the possible differences between disease-inducing strains and non-pathogenic ones. Geographic stratification could in the same way point out disease-inducing mechanisms related to population-specific viral variations.

2. METHODS

2.1. Kit-based target enrichment

Target enrichment is a technique used to improve the ratio of target DNA regions or genomes to the rest of the DNA within a sample. In our first study on HHV-6 (section 4.1), we identified individuals showing strong signs of viral infection, possibly due to congenital integration and its consequential presence of a copy of the virus in each cell acquired in a Mendelian fashion from the parents. Nevertheless, the amount of viral DNA present in a congenitally infected individual is still relatively low, due to the tiny dimension of HHV-6's genome compared to the human's one (160kbp to 3.3Gbp). Sequencing by means of NGS a sample of DNA extracted from a congenitally infected individual would still imply the output of reads mostly belonging to the human host, making the technique expensive when planning to process the 9 samples we had available. In order to produce more virus sequencing data, DNA extracted from LCLs undergo a hybridization-based target enrichment using RNA probes (for details, see section 1.7). The probes were designed and

produced by Roche-Nimbelgen under our guideline of excluding HHV-6's long repeats from the design.

The samples must be prepared for the enrichment by fragmenting the DNA to Illumina-short-read-long fragments (ca. 300 bp in our case) using a sonicator. The fragments will then be processed to become libraries through the addition of sequencing adapters with identification indexes used to separate the reads by sample when multiplexing. The library preparation was done using Illumina commercial kits (Illumina TruSeq® DNA Library Prep LT) on 1µg of sample purified DNA following the supplied protocol, as it was the successive target enrichment.

The results of the capture were positive, as shown in Section 4.1.

2.2. CiHHV-6 identification

In our first HHV-6 study (section 4.1), we sequenced 9 virus genomes in order to study their variability, search eventual signs of geographical stratification, and scan for natural

selection footprints. The samples were DNA derived from LCLs belonging to healthy individuals at the moment of the collection, excluding an acute lytic viral infection. The virus were instead dormant in their latent stage, integrated within the host genome. As explained in section 1.4, HHV-6 seldom integrates within the host germinal line and can be passed to the offspring in the form of CiHHV-6. During the last years, an increasing number of studies focus on the difference in the biology of the active and congenital form of HHV-6, showing evidence that range from different variability rates to different associated pathologies (CIT.), and arising the need to discriminate between them. Our analysis showed congenital infection as the most likely scenario for our samples. Nevertheless we cannot rule out another possibility: the integration of the virus in few B-cell of the host followed by a clonal expansion of the infected cells upon transformation and LCLs establishment. Even if somatic integration occurs in a small number of cells (CIT: Morissette & Flamand 2010) and the probability of clonally amplifying one of those is very small, the samples had to be tested to prove the presence of CiHHV-6.

The more classical methods of detection of CiHHV-6 are composite analysis that uses FISH, a method capable of marking target DNA with fluorescent probes visible when looking at the chromosomes, and amplification of the virus-chromosome junction region for confirmation (CIT.). We choose a easier and faster approach, made popular in the field by Ruth Hall Sedlack and Joshua Hill: measuring the virus/human genome copies through Droplet Digital PCR (ddPCR) (CIT.). Described by Skyes et al. in 1992, ddPCR is a technique that provides ultrasensitive and absolute nucleic acid quantification, designed particularly for targets in complex background or in low abundance. This technique relies in the partition of the PCR volume in a large number of parallel micro-reactions (in the order of tens of thousands) by generating water-in-oil picoliter droplets. The volume is so small to achieve the presence of a single target molecule per droplet, allowing for all-or-nothing results (*i.e*. each droplet will be counted only as positive or negative of amplification). The presence of amplification is measured thanks to florescent

probes added during the reaction to the newly synthetized target copies. Through a Poisson distribution, and thanks to the high number of droplets, and thus reactions, it is possible to estimate with high accuracy the target copy number in the sample (CIT.).

In our HHV-6 study we performed a ddPCR reaction per each sample using a set of two primers-probes: one designed to detect the virus, and one a human single-copy gene. A copy number ratio around 1 would be evidence of the presence of a copy of the virus per cell, a number that is far from the strongly lower one that would be measured for a non-congenital HHV-6 integration (CIT.). Primers and probes sequences are reported in TABXXX.

2.3. Cost-effective virus target enrichment

The study on HHV-6 described in section 4.1 showed interesting yet cryptic results due to the low sample size available. In order to describe the variability and geographical

patterns of HHV-6, we would need to be able to obtain a larger number of genomes from diverse geographical origin. This arose two main questions: 1-Can we sequence latent virus genomes in non-congenital form? The congenital form of HHV-6 affects around 1% of the world population, meaning that to be able to achieve the latent virus genome from individuals presenting non-congenital HHV-6 would improve greatly the available samples, and make easier the collection of new ones. The limitation to the sequencing lies in the extremely low amount of latent non-congenital virus copy in a biological compared to a congenital one, arising doubts on the feasibility of the procedure. Lastly, the possibility to enrich virus genomes in very low copy number would open the possibility to study other Herpesvirus that achieve latency without integration, and thus do not present a congenital form, such as EBV. 2-Can we achieve a method that is cost-effective? To study variability and geographical stratification implies to work on the highest number of sample possible, imposing an economic hard-limit if the production of the virus sequences is too expensive.

The production of virus sequences implicates the development of in-house methods to perform the two principal steps of the procedure: library preparation and target enrichment (results described in section 4.4).

2.3.1. In-house library preparation

The preparation of a sequencing library from DNA is the process of transforming the sample in fragments of homogeneous length bonded at both ends to oligonucleotides needed for sequencing. In order to understand the generation of a library, we need do spend a few words on the structure of a finalized library fragment.

A library fragment is composed of an "insert", or the fragment of sample DNA itself, bounded by the two adapters. These are double stranded oligonucleotides of varying length (depending on the sequencing technologies the library is made for), that are needed for hybridization to the sequencer flow-cell, and therefore for the sequencing step. The adapter can contain a stretch of a few nucleotides (usually 5-7) that are constant within the library of a sample, but different

between samples, making the individual origin of each read recognizable by the sequencer. This allows to sequence multiple samples at once, a process called multiplexing. Sequencing a single sample per run, particularly for short genomes as virus ones, would produce an excessive amount of data, effectively wasting reads and economical resources. The presence of an index sequenced just before each read allows instead to run multiple samples at once, decreasing sensibly the overall cost.

The protocol we used for library preparation was an adaptation made by David A. Hughes of the ones proposed by Meyers & Kircher, and Kircher et al. in 2010 and 2011 respectively (CIT.).

The samples we used for testing were healthy human saliva samples, a tissue chosen for the good balance of easiness of sampling, and high titration of EBV and HHV-6, for both of which the main modality of transmission is through saliva (CIT.). 2 mL of saliva was collected per individual, and DNA

was extracted through a standard absolute precipitation method and purified using 2-fold volume of SPRI beads (see detailed protocol below).

The library preparation begins with the fragmentation of the sample DNA in stretches of a length that depends on the sequencing technology that will be used in the sequencing process. We tested the preparation of libraries meant to be sequenced on Illumina MiSeq system, running the machine to output 150bp for both the forward and reverse strands of each fragment. Being the sequencing pair-end, the overlap of the forward and reverse of each read would make more reliable the mapping process and the building of an assembly, but would lower the amount of information acquired because of redundancy of the bases. The best option in our case would be a fragment distribution peaking at around 300bp.

Fragmentation was at first tested using a Diagenode Bioruptor Plus Ultrasonicator. In order to find the settings most fitting for our experiments, we performed a series of

fragmentations with different number of cycles, and measured the results using an Agilent Bioanalyzer 2100. The samples used for the fragmentation tests were LCL DNA (50µl aliquots containing 300ng of DNA). All samples were fragmented doing 30s ON/ 30s OFF cycles with the instrument set at LOW intensity (160W). The testing was done performing 5, 10, 20 and 30 cycles. Being the results unsatisfactory, we repeated the testing setting the intensity to HIGH (320W). The results (section 4.4)[bioanalyzer at page 4-5 of the 1st lab book] showed variability, probably due to the change in effect the sonication would have with small changes in the conditions of the water bath, leading us to choose instead the Covaris M220 Focused Ultrasonicator, a system with higher result replicability. In order to select the correct settings for the Covaris run, we took the ones supplied with the system itself for 300bp fragments from genomic DNA, and performed a time curve starting from 120 seconds to 200 seconds per cycle, with steps of 20 seconds (Duty cycle=10%, Intensity=5, Cycle/burst=200). The results of the time curve, [to add to results section, in lab book page 15] showed an almost identical fragmentation of the samples (ca 200bp peack)

regardless of the time that each cycle lasted, evidence of lack of over-fragmentation at our settings under that length. We lowered the time of exposure to sonication per cycle, reaching in a final the optimal setting for our experiments: 80 seconds.

The protocol consisted in the preparation of the hybridization of the P5 and P7 Illumina adapters (adapter sequences in Supplementary Table X)[generate tab], and the library preparation itself. This comprised 4 stages, shown in Figure X [add Meyer et al figure]: 1- Blunt ends repair: fragmented samples often break in uneven position in the 5' and 3' strands due to the mechanical nature of the sonication. In order to ligate the adapters to the inserts, the ends must be even, which is achieved in this step. 2- Adapter ligation: The hybridized pre-adapters (first part of the complete adapters, used as template for the index addition and completion of the sequencing adapter) ligated to the ends of the inserts. 3- Adapter fill-in: The short IS3 adapter is elongated using as template IS1 on the 5' end and IS2 on the 3' end (Adapter

sequences reported in Supplementary Table X) [add sequences to tab]. 4- Indexing PCR: The filled adapters are used as template for a PCR reaction in which an index would be inserted at the end of the adapters, followed by the rest of the sequencing adapters. To notice that this protocol allows to add different indexes at the two endings of the fragment, actively allowing for high multiplexing thanks to the combination of the two indexes for sample assignation. Also, in order to further decrease the costs, we produced the SPRI beads ourselves following the protocol published by Rholand et al. [add Nadin's paper citation].

The protocol proceeds as follows:

Starting Material:

- 500-1000 ng dsDNA
- The amount of starting material should be chosen to insure that the representation of the target molecules in the final library is sufficient
- Final yield of the library is about 10% - 20% of starting material

Identify indexes:

- Identify what subset of indexes you will use to tag your libraries. You must insure equal base (balanced) composition across indexes to insure no difficulties during the image analysis of the sequencing by synthesis steps on the illumina system (CIT.)

Prepare Buffers:

- EBT (10mM Tris-HCl pH 8-8.5; 0.05% Tween-20)
- Oligo hybridization buffer (10mM Tris-Cl; 1mM EDTA; 500 mM NaCl; pH 8)

Scheme in Short:

1. Prepare adapters
2. Blunt End Repair
3. SPRI Bead Clean-up 1
4. Adapter Ligation
5. SPRI Bead Clean-up 2
6. Adapter Fill-In
7. SPRI Bead Clean-up 3
8. Indexing PCR

9. SPRI Bead Clean-up 4

1- Prepare adapters:

This will produce enough adapter mix for 200 reactions. **Store at -20°C.**

Assemble the following two reactions:

I. Hybridization mix for adapter P5 (200µM)

Reagent	Stock [ ]	Final [ ]	Vol (µl)
IS1_adapter_P5.F	500µM	200µM	40
IS3_adapter_P5+P7.R	500µM	200µM	40
Oligo hybridization buffer	10x	1x	10
H_2O	NA	NA	10

II. Hybridization mix for adapter P7 (200µM)

Reagent	Stock [ ]	Final [ ]	Vol (µl)
IS1_adapter_P7.F	500µM	200µM	40

IS3_adapter_P5+P7.R	500µM	200µM	40
Oligo hybridization buffer	10x	1x	10
ddH_2O	NA	NA	10

1. Mix the reactions by pipetting up and down at least 3x.
2. Incubate in thermocycler for:

Step	Temp	Time	Cycle
Denature	95°C	10sec	1
Hybridize	95°C - 12°C	0.1°C/sec	RAMP
Hold	8°C	forever	1

3. Combine both reactions into a single tube
    a. You now have a ready to use mix (100µM / adapter)

2- Blunt End Repair

1. Add Positive Control ad negative control (50µl ddH_2O) to 96-well plate
2. Prepare the following master mix

Reagent	Stock [ ]	Final [ ]	Vol. (µl) / rxn
ddH_2O	NA	NA	7.12
Buffer Tango	10x	1x	7
dNTPs	25mM each	100uM each	0.28
ATP	100mM	1mM	0.7
T4 PNK	10 U/ul	35 U	3.5
T4 DNA Poly.	5 U/ul	7 U	1.4

3. Mix carefully by flicking tube or stirring and pipetting up and down carefully.
4. Add 20ul of master mix to each 50ul sample.
5. Incubate in thermocycler:

Step	Temp	Time	Cycle
Incubate	25°C	15mins	1
Incubate	12°C	5 mins	1
Hold	8°C	forever	1

6. Place on ice or proceed immediately to next step

3- SPRI Bead Clean-up 1

1. Add 154 µl of beads to the Blunt End Repair product
    a. That is a 2.2-fold volume of beads to reaction volume
2. Gently pipette up and down 10x
3. Incubate at RT for 5 min
4. Place on magnetic stand and allow to stand for ~ 2min
5. Remove supernatant without disrupting beads
6. Wash with 190 µl of FRESHLY PREPARED 70% EtOH – ALWAYS leaving plate on stand – let sit for 30 secs and remove EtOH

7. Leave plate on magnet and Repeat Wash (step 6)
8. Remove residual ethanol with pipette, carefully
9. Air Dry for 10min
10. Add **20 µl** of EBT
11. Remove from magnet
12. Resuspend by vortexing repeatedly
13. Spin down very briefly to collect volume at bottom of plate
14. Place back on Magnetic Stand, let stand for 1-2mins, and remove products (supernatant) and add to a new 96-well plate

4- Adapter Ligation

1. Prepare the following 40 µl reaction

   NOTE: If 5x ligase buffer contains a precipitate after thawing, warm the buffer and vortex until precipitate dissolved. Remember that the buffer **contains ATP**, which is susceptible to thawing cycles.

Reagent	Stock [ ]	Final [ ]	Vol. (µl) / rxn
ddH_2O	NA	NA	10
T4 DNA ligase buffer	5x	1x	8
Adapter Mix	100µM each	2.5 µM each	1
T4 DNA ligase	5U / µl	5U	1

2. Mix carefully by flicking tube or stirring and pipetting up and down carefully. **DO NOT VORTEX !!**
3. Add 20µl of master mix to each sample
4. Mix by pipetting gently
5. Incubate in thermocycler:

Step	Temp	Time	Cycle
Incubate	22°C	70mins	1
Hold	8°C	forever	1

5- SPRI Bead Clean-up 2

Repeat SPRI Bead Clean-up 1 using 88µl of beads per sample (2.2-fold beads volume per reaction)

6- Adapter Fill-In

1. Prepare the following 40 µl reaction

Reagent	Stock [ ]	Final [ ]	Vol. (µl) / rxn
ddH_2O	NA	NA	14.1
Bsm buffer	10x	1x	4
dNTPs	25mM each	250µM each	0.4
Bsm polymerase, large fragment	8U/ul	12U	1.5

2. Add 20µl of master mix to each sample
3. Mix well, **DO NOT VORTEX**

4. Incubate in thermocylcer:

Step	Temp	Time	Cycle
Incubate	37°C	25mins	1
Hold	8°C	forever	1

7- SPRI Bead Clean-up 3

Repeat SPRI Bead Clean-up 2.

8- Indexing PCR

1. Prepare the following 50 µl PCR reaction
    a. Dispense the master mix into a 96-well plate
    b. Add a DIFFERENT SET of Indexing Primers to each well
    c. Add the template DNA

Reagent	Stock [ ]	Final [ ]	Vol. (µl) / rxn
ddH_2O	NA	NA	21.5 µl

AccuPrime reaction mix	10x	1x	5 µl
AccuPrime Pfx polymerase	2.5U / µl	1.25 U	0.5 µl

2. Add the indexing Primers

Reagent	Stock [ ]	Final [ ]	Vol. (µl) / rxn
P5 index primer	10µM	300nM	1.5µl
P7 index primer	10µM	300nM	1.5µl

3. Add the 20µl of template DNA

4. Place in a thermocycler for PCR amplification

Step	Temp	Time	Cycle
Denature	95°C	2 mins	1
Denature	95°C	15 secs	10x
Anneal	60°C	20 secs	10x

Extend	68°C	30 secs	10x
Hold	8°C	forever	1

9- SPRI Bead Clean-up 4

Repeat SPRI Bead Clean-up 1 using 110µl of beads per sample (2.2-fold beads volume per reaction), and resuspend in 25µl of EBT.

Library confirmation

The library preparation included a positive control as sample, represented by a PCR product of a length similar to the peak length of the sample fragments distribution (330bp in our case). Three aliquots of the positive control were retrieved before the library preparation, after the adapter ligation and fill in, after the indexing PCR. Because of the length of the P5 and P7 at first, and of the index and rest of the sequencing adapter at the end, the three aliquots would show a fragment length of 330bp, 390bp and 420bp respectively, indicating the

success of the library preparation [insert positive control bioanalyzer in results].

2.3.2. Target enrichment

The target enrichment was performed using commercial baits, and a set of 14 libraries built following our in-house protocol where produced starting from saliva samples collected from different individuals. All samples were tested for viral presence (EBV) through real time PCR on a Thermo Fisher(10 µl reaction composed of 3µl of 5X HOT FIREpol EvaGreen qPCR Supermix (Solis Biodyne), 0.5µl of each primer solution at 10µM, 5µl of water and 1µl with 20ng of sample DNA. Primer sequences reported in Supplementary TabX [add outer inception primer sequences]. The reaction settings were: 50°C 2' -> 95°C 12' -> 40 cycles of 95°C 15'', 58°C 60'', 72°C 30'' ->72°C 30'').

The target enrichment baits were designed and produced by MYcroarray under our guidelines. The design covers the whole genome of 3 reference virus sequences, excluding

their main annotated repeated regions: EBV type 1 (NC_007605), HHV-6A (NC_001664) and HHV-6B (NC_000898) (Supplementary table X). [add main repeats coordinates]. The baits were RNA 120mers designed to have a 4X tiling on the target. The protocol followed the guidelines supplied by MYcroarray together with the baits, with two differences: 1-The post-capture washing were performed at 68°C instead of 65°C in order to improve stringency and lessen the formation of daisy-chains due to the complementarity of the adapters (CIT.). 2-The blocking oligos used were only 2 instead of four, one complementary to the adapter of one strand of the 5' fragment end, and the other to the adapter of the opposite strand at the 3' fragment end. This method would insure blocking of the adapters, and the absence of blocking oligo dimers (being the two blocking oligos not complementary). In this fashion we are able to control with precision the concentration of the blocking oligos, usually altered by random dimers formation (CIT.).

The enrichment was tested using relative real-time PCR on the pre- and post- enrichment sample pools, testing for the amplification of EBV, HHV-6, and the Human DNA content.

The final enriched samples pool was sequenced using an Illumina MiSeq System on a 150X2 pair end run.

2.4. EBV-immortalized cell cultures and viral load confirmation

The GWAS project described in section 4.2 involved the identification of region of the host genome associated to the viral load of EBV within the LCLs. In order to be used as a phenotype trait in the GWAS, the viral load in the LCLs must be shown to be stable in time. If it varied, the viral load values we estimated would only show the virus copy number at the time of measurement, and would thus be unsuited for the study. In order to demonstrate the stability of the trait in time, we cultured 7 LCLs generated from biological samples extracted from random individuals of the study data set. The samples belonged to individuals present in the 1000 Genome

Project under the identification IDs: HG01277, HG00245, HG00362, HG00657, NA18999, NA18502, NA19382.

Each cell culture was divided in 3 replicates, and each replicate cultured at the same condition for 6 passages (3-4 days between passages), keeping the cell concentration above 200K cells/mL and under 2M cells/mL (for culture method details, see attached paper in section 4.2).

The relative viral load between LCLs and time points (*i.e.* passages) was calculated through real time PCR, using a primer-probe set ("Eagle" set, sequence in Supplementary TableX[add sequence to tab]) designed to fall on a region of the IR-1 repeat. The IR-1 repeat is a repetitive element that varies in copy number between EBV strains, but is stable in the transforming strain at ca. 8 copies (CIT.). The transforming EBV strain used to immortalize the LCLs is the one for which we are measuring the viral load, allowing us to amplify a region that is 8 times more abundant than a single-copy region, and increasing our real-time PCR assay sensibility.

The measurements of copies per cell of EBV were estimated through an *in silico* method based on the ratio of depth of coverage of EBV to the depth of coverage of the human genome, taking in account the diploidy of the human genome. In order to validate this method, we performed a real-time PCR on DNA belonging to 13 LCLs generated from biological samples extracted from random individuals of the study data set, and compared the results with the *in silico* estimation of the same individuals. The positive correlation was remarkably high (see section 4.2), validating the method.

2.5. EBV genes amplification and sequencing

The geographic patterns of variability of EBV are a key factor to understand the genetics of this virus, especially because of the strong confounding effect that these can have on the interpretation of disease-association studies. In order to explore the stratification of EBV, we aimed to sequence and compare a cryptic set of genes, mainly related to the virus latency, derived from healthy individuals of diverse

geographical origin (Table X) [table with amplicons, genes, extension times and sensibility in copy number]. The data set we worked on, kindly provided by Mark Stoneking from the Max Planck Institute for Evolutionary Anthropology, was composed of DNA extracted from saliva of 354 healthy individuals at the time of sampling, from 11 countries around the world (see Supplementary material Table X)[Add sample list with origins].

The project aimed to obtain the sequences of the selected set of genes by means of long range PCRs, followed by transformation of the amplicons in sequencing libraries tagged for multiplexing, and sequencing using Illumina short-read technology.

The first step has been the design of primer sets able to amplify the selected set of genes. After a trial-and-error stage, we ended up with a set of primers of 14 pairs, 2 of which showed the same yield at the same reaction condition, allowing us to use them in the same amplification reaction. Primer sequences are reported in Supplementary Table X.

[same table as before]. All primer tests, and successive amplification have been performed using NEB LongAmp Taq PCR kits, following the supplied guidelines and setting the annealing temperature at 63 °C. Extension time depended on the length of the amplicons, and are listed in Supplemntary Table X.[again, same table as before].

It is known that the EBV copy number in saliva depends on the individual, and on the time of collection, and is a factor that varies greatly (CIT). Because of this, in order to avoid performing reactions on samples with low presence of the virus, the selected primers were tested on samples with increasingly lower EBV copy number in order to evaluate the sensibility of the reactions in the amplification of the targets (12000, 6000, 3000, 1500, 300 virus copies per reaction). The primers pairs were capable of detecting the targets down to 500-1500 copies of the virus per 200ng of total DNA (Figure X) [Gel of the last primer tests]. The virus copy number was measured in the samples used for testing the primers through real time PCR, using a series of standards. The standard were generated using a pair of primers

designed to amplify a 243bp region of the BBLF4 gene ("Banjo" amplicon, sequence in Supplementary Tab X)[add seq to table]. The amplified Banjo amplicons were purified using a QIAquick purification kit, and quantified using an Invitrogen Qubit system. The concentration of the number of amplicons was calculated dividing the measured DNA concentration for the weight of a single copy of the amplicons. The stock Banjo amplicons solution was diluted 1/25 7 times, generating solution ranging from 17B to 100 copies of the amplicon per µl. We then designed a second primer pair, nested within the Banjo amplicon ("Kazooie" primers, sequence in Supplementary Tab X)[add seq to table]), that would be used for the real time PCRs used for determining the copies of EBV present in the samples. Being the Kazooie amplicon able to amplify a region of the Banjo amplicons within the standards, as well as the actual EBV within the samples, they can be used for absolute quantification purposes. The pipetting error of 7 1/25 dilutions can get to a magnitude that do not allow us to give exact copy number values, but the results can be used to have a relatively

precise range, which is enough for the purpose of selecting the samples for testing.

The geographical panel was composed of DNA samples extracted from saliva, and DNA concentration was measured with Invitrogen Qubit fluorimeter. The amount of total DNA varied from 200ng to >50µg depending on the sample, leading to the exclusion of 63 samples due to the relatively high quantities needed for the following experiments (200ng of DNA per reaction, 10 reaction minimum per sample). This filter left us with 291 samples.

The number of individuals that could undergo the amplification of the selected genes was limited by time and budget. We had to select the ones with the higher probability of success during the experiments, which translates to the ones with higher EBV copy number. Lacking the availability of the thermocycler system to perform qPCR to measure the viral load, we tested the amplification of the amplicons that showed higher sensibility during the primer testing phase on all the samples data set (Primers sets 1 and 8, see

Supplementary Table 2). We selected the 120 individuals with higher EBV copy number, which constituted our final data set for the long range PCRs.

The selected genes were amplified using our 14 primers pair set on the chosen 120 individuals, and successful amplification was verified on a 1.5% agarose gel together with NYZDNA ladder III. 17 individuals showed less than 3 of the 14 total amplicons amplified, leading to their exclusion from the study.

All PCR reactions were treated with 5µl of 5mg/mL of Qiagen RNAse A solution at 37°C for 35', and purified using 1.5X of SPRI beads (see in-house library preparation protocol, section 2.3.1). The final products were quantified on Thermo Fischer Nanodrop 2000 system, and the amplicons were equimolarly pooled per each individual.

Sequencing libraries were built starting from 500ng of each individual amplicons pool following the in-house protocol described in section 2.3.1, using double indexing. An equimolar pooling of the libraries was prepared, and sent to sequence on an 150X2 Illumina MiSeq System run.

3. OBJECTIVES

This thesis explores virus genetics at a population level, from the generation of data to the analysis of the variability specifically focusing on two cryptic virus species: Human herpesvirus 4 and 6.

The following specific objectives have been addressed:

1- Exploration of cost-effective methods to generate virus genetic data in healthy individuals.

1.1- Commercial target enrichment of Human herpesvirus 6 in congenital state from 1000 Genome Project individuals.

1.2- Design of long range PCR assays to generate gene data from Human herpesvirus 4 from host of diverse world-wide origin.

1.3- Adaptation and testing of custom target in-house enrichment protocols to capture Herpesviruses in latency in a cost-effective manner.

2- Analysis of virus variability in evolutionary and geographic frameworks.

2.1- Whole genome comparative analysis of the congenital form of Human herpesvirus 6 from healthy hosts with diverse geographical origin.

2.2- Stratification analysis of Human herpesvirus 4 in a large geographical panel using latency genes.

2.3- Analysis of the relationship between Human herpesvirus 4 infection and the genetic architecture of the human host.

[Pàgina en blanc]

4. RESULTS

4.1. Whole genome diversity of HHV-6 derived from healthy individuals of different geographical origin

Telford, M., Navarro, A., Santpere, G. Whole genome diversity of HHV-6 derived from healthy individuals of different geographical origin. *In submission*

Telford M, Navarro A, Santpere G. Whole genome diversity of inherited chromosomally integrated HHV-6 derived from healthy individuals of diverse geographic origin. Sci Rep. 2018;8(1):3472. DOI: 10.1038/s41598-018-21645-x

4.2. Genetic factors affecting EBV copy number in lymphoblastoid cell lines derived from the 1000 Genome Project samples

Mandage*, R., Telford*, M., Rodríguez, J.A., Farré, X., Layouni, H., Marigota, U.M., Cundiff, C., Heredia-Genestar, J.M., Navarro, A., Santpere, G. Genetic factors affecting EBV copy number in lymphoblastoid cell lines derived from the 1000 Genome Project samples. (2017) PLoS ONE 12(6):e0179446.
https://doi.org/10.1371/journal. pone.0179446

Mandage R, Telford M, Rodríguez JA, Farré X, Layouni H, Marigorta UM, et al. Genetic factors affecting EBV copy number in lymphoblastoid cell lines derived from the 1000 Genome Project samples. PLoS One. 2017 Jun 27;12(6):e0179446. DOI: 10.1371/journal.pone.0179446

4.3. EBV latency gene stratification

The geographic patterns of variability of EBV are a key factor to understand the genetics of this virus, especially because population stratification is a major confounding effect on disease-association studies. In order to explore the stratification of EBV, we aimed to sequence and compare a cryptic set of genes, mainly related to the virus latency, derived from healthy individuals of diverse geographical origin.

The selected genes to amplify, listed in Table 5, were enclosed in a set of 10 amplicons with lengths varying from 1000-6500bp. Different sets of similar primers were designed per each amplicon, in order to identify the pair more suited for amplification in our complex samples (i.e. samples with very high background DNA compared to the target one).

Amaplicon ID	CDS	Approx. Length (bp)
1	LMP2	1200
4	S/T kinase, BDLF4	3300-4200
5	EBNA3C, BZLF1-BRLF1 , EBNA1	2700-3700
B	LMP2, EBERs	2900
D	EBNA3A, EBNA3B	6400
6	LMP1	3300
7	BHRF1	3400
8	LMP2	1900
3-5	EBNA2	2200
3-9	EBNA3C	3500

Table 5. Amplicons sets with respective amplified coding region and amplicon length.

The primers were tested for optimum annealing temperature on samples with higher EBV titration. The samples' DNA were extracted from LCLs immortalized with EBV, and therefore presented variable, but very high EBV copy number. The measure in two samples of EBV-immortalized Lymphoblastoid Cell Lines (LCLs) of the viral transforming strain (B95-8) copy number compared to the natural infecting strain showed a ratio of thousands of folds higher (Telford, Santpere & Navarro, unpublished data). The primers, which we calculated to have optimum melting temperature at 61-

62°C, were tested repeating the amplification with different annealing temperature, ranging from 59-67°C. The test was repeated with three different commercial kits: PrimeSTAR® GXL DNA polymerase (Takara), LongAmp® Taq PCR kit (NEB), and ALLin™ RPH polymerase (HighQu). While the Allin RPH kit gave little amplification in all reactions, the PrimeStar GXP DNA polymerase and LongAmp taq PCR kit showed good amplification in all reactions, with the latter producing a more intense band in the electrophoresis gel. Because of the increased amplification and the lower cost, we choose NEB LongAmp kit for the successive long range PCR.

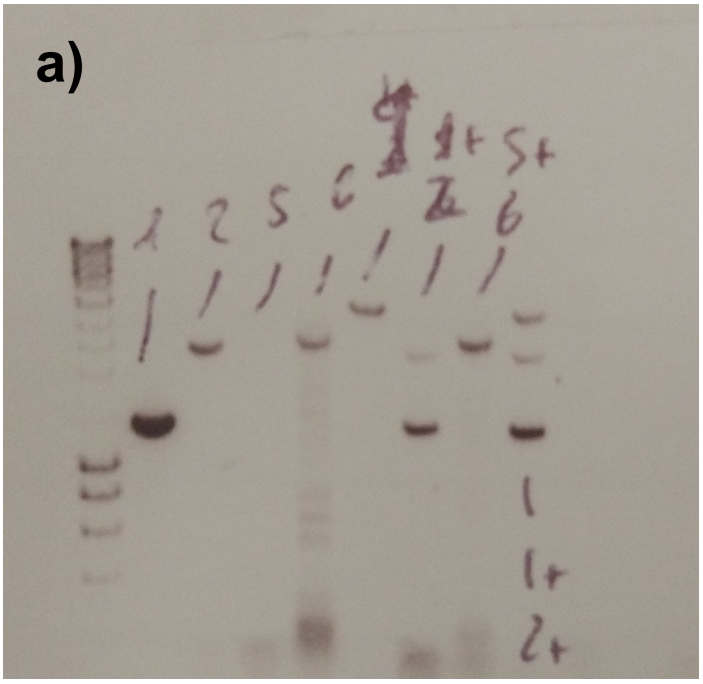
A second round of testing involved the amplification of multiple amplicons in a single reaction, in order to decrease the number of PCR to perform on each individual. The amplification of multiple targets was successful, but in most cases the amount of each target was different within a same reaction, as shown in the example in Figure 14a. This would hinder the amplicons equimolar pooling per individual previous to sequencing, leading to an eventual strongly heterogeneous coverage. Two sets of amplicons were

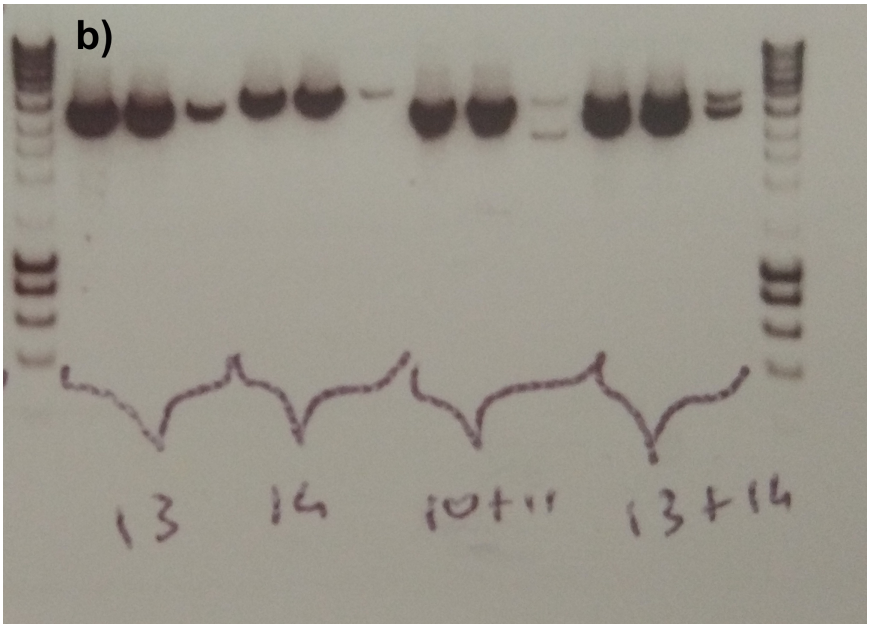
amplified in similar fashion when targeted in the same reaction: 10 + 11, and 13 + 14. A second test was performed on these two primers sets to prove their similar amplification in a single reaction by using three samples with decreasing EBV copy number: 6000, 1500, 500 copies per reaction (Figure 14b). While at high copy number the band of the amplicons were too intense and merged in a single one, when using low EBV copy number the amplification appears clearly similar for both amplicons in each set. Each of these two sets were used in a single reaction during the project.

It is known that the EBV copy number in saliva depends on the individual, and on the time of collection, and is a factor that varies greatly(Ling et al., 2003; Q. Y. Yao & Rickinson, 1985). Because of this, in order to avoid performing reactions on samples with low presence of the virus, the selected primers were tested on samples with decreasing EBV copy number in order to evaluate the sensibility of the reactions in the amplification of the targets (12000, 6000, 3000, 1500, 500 virus copies per reaction). The sensitivity of each primer set ranged between 500-1000 copies. In Figure 14c an example

of positive results (amplicon B test) show amplification down to the minimum of 500 copies.

Overall, we achieved a set of solid amplicons with high sensitivity fit for our purposes.




b)
13
14
10+11
13+14

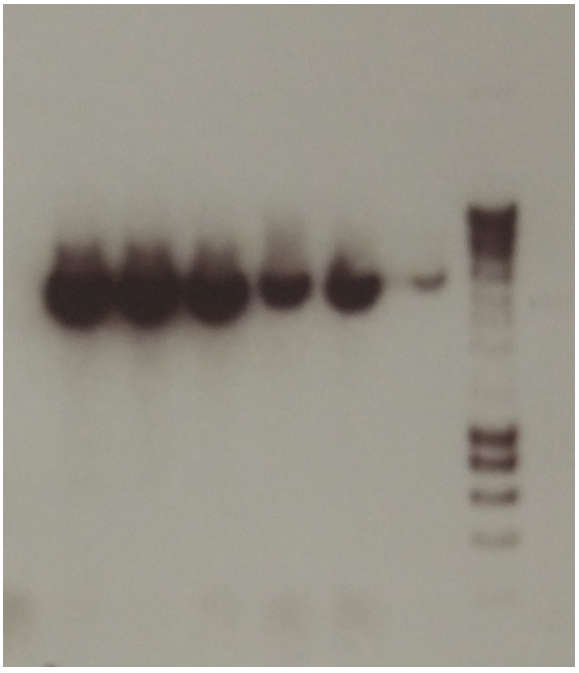


Figure 14. Multiple amplification per reaction test. The written number indicates the amplified amplicons. a) Unsuccessful tests example. The amplicons were amplified singularly (left), and multiplexed (right). b) Successful test example. Amplicons 10+11 and 13+14 were amplified three times on samples with decreasing EBV copy number. In the right-most reaction of each set of amplicons the bands of both amplicon appear at the correct fragment length, and with similar intensity. c) B amplicon test on (right to left) 12000, 9000, 6000, 3000, 1500, 500 EBV copies per reaction.

The total sample panel was composed of 353 saliva extractions. A filtering of the samples was done excluding the ones with too little total DNA for the following amplifications, leading to the removal of 62 extractions. A second filter was done performing the amplification in all individuals with the most sensible sets of primers (1 and 8) in order to identify the samples with too low EBV copy number for successful long range PCRs. This lead to the removal of an additional 171 individuals. Lastly, the amplification of more than 7 of the 10 amplicons on 18 individuals prompt their exclusion from library preparation and sequencing. The final data set constituted 102 samples from 11 countries (Table 6).

Even after the exclusion of many samples from the original panel, the final data set is large and diverse, covering many origin location and the main continents, and making it suited for a geographical stratification analysis.

Continent	Country	N. of samples	Final data set samples
Asia	China	14	4
Africa	Congo	49	12
	South Africa	10	4
Eurasia	Georgia	95	30
Europe	Germany	10	1
	Poland	30	10
	Turkey	50	16
North Ameri	California	25	3
	Louisiana	25	5
South Ameri	Argentina	20	6
	Bolivia	25	11
	TOTAL	353	102

Table 6. Sample number per country and continent of origin in the whole sample panel, and that filtered, that has been used as data set for the study.

The long-range PCR products that followed were run on electrophoretic gel for confirmation, purified using the SPRI beads method and quantified for DNA amount. The quantification and amplicon weights were used to pool

equimolarly all amplicons per individual. The resulting 102 pools were converted to sequencing libraries following the B.E.S.T. protocol(Caroe et al., 2017), and labelling each one with a unique set of 2 indexes for multiplexing. The library preparation was followed by equimolar pooling of the libraries in a single pool, and sequencing on a single flow-cell lane.

The sequenced pool's reads were mapped against the EBV reference sequence, and depth of coverage per each coding region was calculated (Table 7).

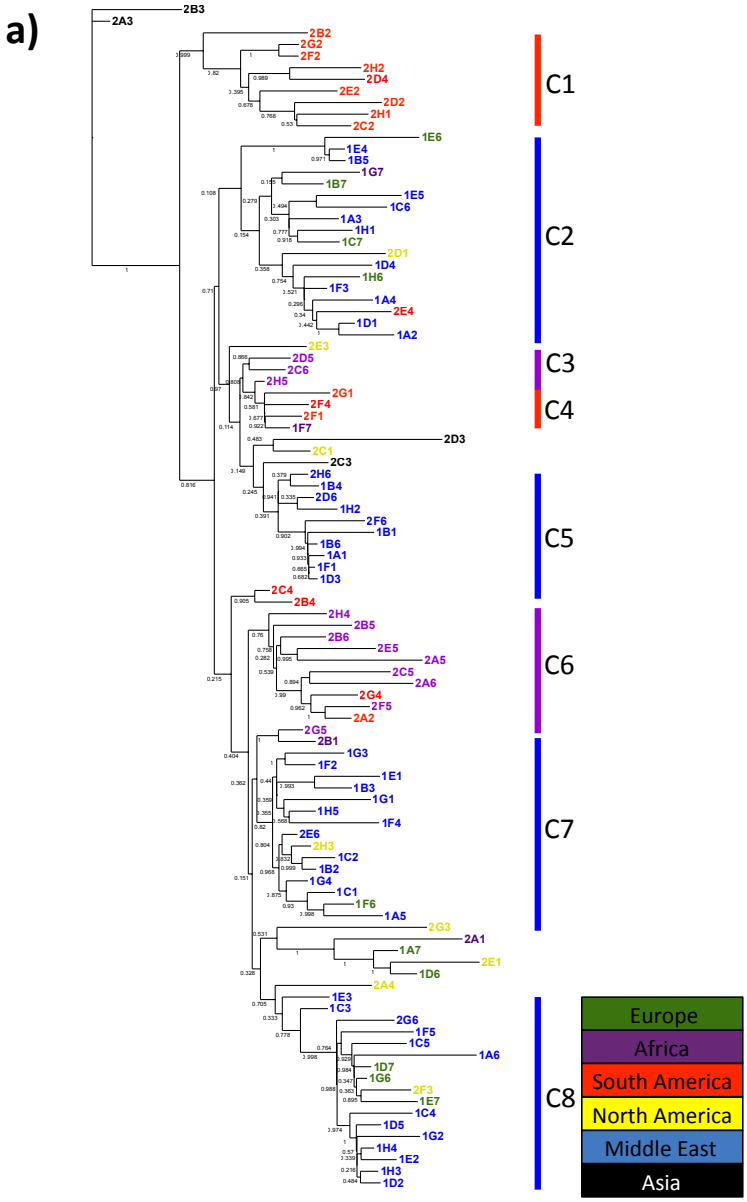
CDS	Avg coverage	St. Deviation	Discarded individuals
EBNA2	1	2	71
EBER2	9	7	24
EBNA3C	36	57	36
EBNA3B	48	52	28
EBNA1	53	28	1
EBNA3A	61	63	32
LMP1	76	43	0
LMP2B	85	62	4
BRLF1	98	51	2
BZLF1	111	62	3
BHRF1	116	69	0
BDLF4	125	61	3
EBER1	127	84	6
BGLF4	141	67	1
LMP2A	179	74	0

Table 7. Average depth of coverage and standard deviation per coding region. The numbers of individuals excluded for

having a depth of coverage lower than 2 are listed in the right-most column.

The analysis of the sequenced data show heterogeneous coverage per coding region and per individual, but overall the coverage is enough for the following analysis. EBNAs (EBNA1, EBNA2, EBNA3A-C) appear to be the ones with lower coverage, with EBNA2 being the one with the lowest one and highest number of individuals showing average depth of coverage lower than 2 (71 discarded individuals). The sequences are meant for comparative analysis of the genes in order to study genetic distances between individuals, and while it is possible to estimate missing data, the 31 individuals passing quality filters for EBNA2 are not enough to correctly perform the estimation. This lead to the exclusion of this gene from the data set in the supertree analysis. The other genes resulted in high quality sequences in almost all individuals, with the exception of EBER2 and EBNA3 that still had enough individuals for tree coalescence analysis.

Two main phylogenies were constructed to represent the relationships between the data set taxa: collapsed data tree and supertree. While the first was built collapsing together all the data per individual as a single sequence (Figure 3a), the second took advantage of an algorithm that coalesce all the single gene trees in a consensus supertree (Figure 3b).


a)
2B3
2A3
2B2
2G2
2F2
2H2
2D4
2E2
2D2
2H1
2C2
1E6
1E4
1B5
1G7
1B7
1E5
1C6
1A3
1H1
1C7
2D1
1D4
1H6
1F3
1A4
2E4
1D1
1A2
2E3
2D5
2C6
2H5
2G1
2F4
2F1
1F7
2D3
2C1
2C3
2H6
1B4
2D6
1H2
2F6
1B1
1B6
1A1
1F1
1D3
2C4
2B4
2H4
2B5
2B6
2E5
2A5
2C5
2A6
2G4
2F5
2A2
2G5
2B1
1G3
1F2
1E1
1B3
1G1
1H5
1F4
2E6
2H3
1C2
1B2
1G4
1C1
1F6
1A5
2G3
2A1
1A7
2E1
1D6
2A4
1E3
1C3
2G6
1F5
1C5
1A6
1D7
1G6
2F3
1E7
1C4
1D5
1G2
1H4
1E2
1H3
1D2
C1
C2
C3
C4
C5
C6
C7
C8
Europe
Africa
South America
North America
Middle East
Asia

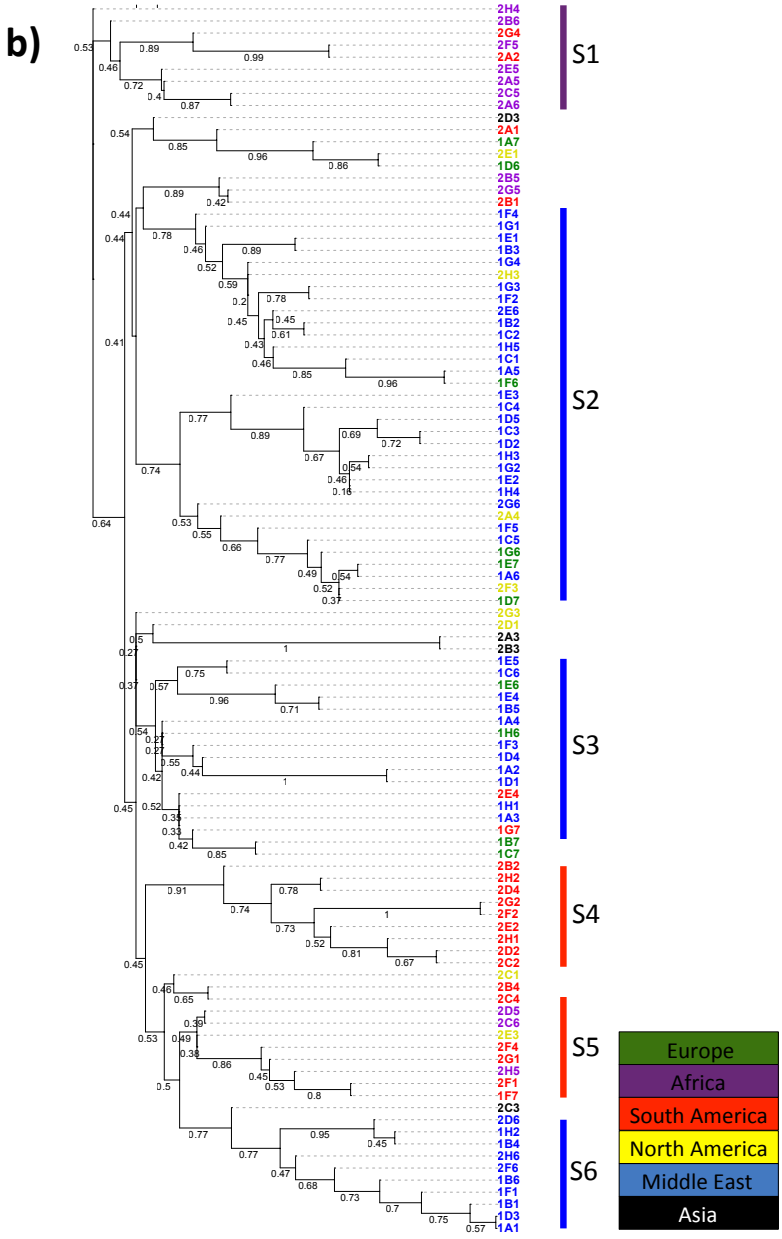


Figure 3. Phylogenetic tree built on the whole data set excluding EBNA2. The strains are coloured based on their origin. The dominating geographical component in each

cluster, when clear, is shown by coloured vertical lines. a) Tree built from the collapsed data per each individual. b) Consensus supertree of the single gene trees. Clusters ID are shown as Cx for the collapsed tree, and Sx in the supertree.

While showing small differences, a background structure appears in both phylogenies. African individuals presents in both trees a strong cluster composed of Congolese strains, a cluster that is basal to the tree in the supertree analysis (cluster C6 and S1). The rest of the African strains derived from South Africa, and cluster together with South American sequences. The low divergence between South African individuals and South Americans remains clear in both phylogenies. The South American sequences present two distinct patterns: the formation of a solid cluster (C1, S4), which is composed exclusively of Bolivian strains, and the admixture of the other ones to African sequences. The Middle Eastern individuals do form aggregations, but in multiple clusters. While in the collapsed tree these cluster appears more admixed with strains of different origins, with the exception of the "purer" cluster C5, in the supertree the

Middle Eastern strain result more clearly separated from the rest of the sequences, resulting in three well-defined clusters (S2, S3, S6).

Overall, the general phylogenies appear influenced by the geographical origin of the strains, even though the stratification signal is likely blurred by the pervasive recombination characteristic of this virus. More detailed analysis of the recombination patterns and population definition, such as Matteo *et al*.'s ones(Matteo et al., 2016), are due to give clarity to the actual patterns presents in out data.

4.4. Towards a cost-effective large-scale target enrichment

To understand organisms such as EBV and HHV-6, their variability and stratification, and their evolution, is a complex challenge. As seen in the study on HHV-6 described in section 4.1, the results are suggestive but yet obscured due to the still low sample size available. The lack of large numbers of complete available genomes, and the lack of diversity in their origin is the main limitation to the understanding of the variability of these viruses. While during the last years this number increased from the order of ten to the order of hundreds thanks specially to Palser *et al.* 2015(Palser et al., 2015) for EBV, the diversity in geographical origin and viral form is still low. The understanding of variability must include the analysis of geographical population structure of the viruses, particularly when trying to link genetic variation to diseases that show solid stratification patterns. Also, EBV and HHV-6 have multiple forms, of which differences are yet not fully elucidated, and should then be studied in a comparative way.

EBV has an acute and a latent form, and it has been studied almost always from lytic-infected samples, limiting potential samples and tissues for collection. Unfortunately, while it is easy to sequence EBV in its acute form because of the high viral load typical of this life-cycle phase, the latency implies such low viral copy number to pose a serious challenge to the production of sequences. The same can be said for HHV-6, but with an additional form of the virus to add to the complexity: the congenital virus. This form, while being easier to sequence than the non-congenital integrated virus because of the copy of the virome present in every cell of the host(Arbuckle et al., 2010; Luppi et al., 1993), impose the need of a third cohort of sequences in order to correctly asses variability. While recent studies are showing the genetic separation between iciIHHV-6 and acute HHV-6(J Tweedy et al., 2015; Joshua Tweedy et al., 2016), the genetic separation between the different forms of the virus is still poorly understood.

The need of larger number of genomes from diverse geographical origin of EBV and HHV-6 presents two main questions: 1-Can latent virus genomes be sequenced in non-congenital form? The congenital form of HHV-6 affects around 1% of the world population, meaning that to be able to achieve the latent virus genome from individuals presenting non-congenital HHV-6 would improve greatly the available samples, and make easier the collection of new ones. The limitation to the sequencing lies in the extremely low amount of latent non-congenital virus copy in a biological compared to a congenital one, arising doubts on the feasibility of the procedure. Lastly, the possibility to enrich virus genomes in very low copy number would open the possibility to study other herpesviruses that achieve latency without integration, and thus do not present a congenital form, such as EBV. 2-Is it possible to develop a cost-effective method? To study variability and geographical stratification implies to work on the highest number of samples possible, imposing an economic hard-limit if the production if the virus sequences is too expensive

It is possible to produce latent virus sequences with the present target enrichment methodologies in latent EBV and in HHV-6 (at least in its congenital form), but it is expensive when trying to generate the large data sets required for a high-resolution variability analysis. In order to tackle this problem, we started the adaptation of in-house cost-effective library preparation and target enrichment protocols aimed at viruses.

4.4.1. Library preparation

The first step was to generate the Solid Phase Reversible Immobilization (SPRI) beads following Rholand and Reich's protocol(Rohland & Reich, 2012). The beads were tested at different beads/sample volume ratios against the Agentcourt® AMPure® X commerical ones. A testing sample batch was prepared with 9 µg of GeneRuler Ultra Low Range DNA Ladder diluted in 180 µl of water. The mix was divided in nine 20 µl samples and purified (1 ladder µg per sample). The purified products were run on a 2% agarose electrophoretic

gel (Figure 15). The freshly prepared beads appeared to work well, with comparable results with the commercial ones.

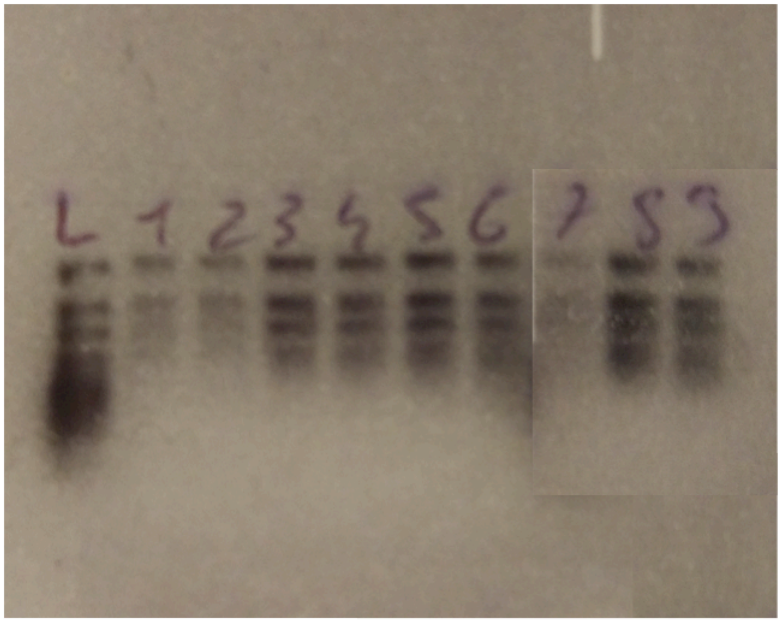
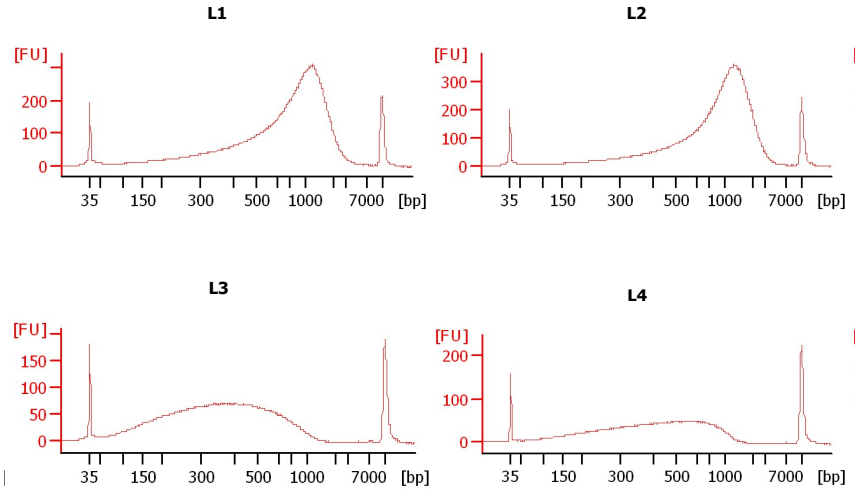


Figure 15. SPRI beads testing. The beads/sample volume ratio, and beads source were: Home-made beads purifications, 1-> 1.3X, 2-> 1.7X, 3-> 2.1X, 4-> 2.3X, 5-> 2.5X, 6-> 3.0X. Agencourt AMPure beads 7-> 1.3X, 8-> 2.1X, 9-> 2.5X.

In order to produce libraries to sequence on an Illumina system, the samples had to be fragmented to a short-read-long length. The length we tested for was a distribution peaking at 300-350bp, in order to sequence the libraries on a

Miseq 150X2 run. Fragmentation was tested performing a series of sonications on identical samples using the same settings but with variable the number of sonication cycles on a Diogenode Bioruptor Plus Ultrasonicator. The settings were an adaptation of Meyer & Kircher's protocol(Meyer & Kircher, 2010), with the number of sonication cycles varying form 5 to 30 and the intensity set at "LOW" (160W). The experiments were repeated in two sets of samples, the first containing LCL DNA, and the second human saliva DNA. Shown in Figure 16, are the fragment length distributions measured through Agilent Bioanalyzer 2100.



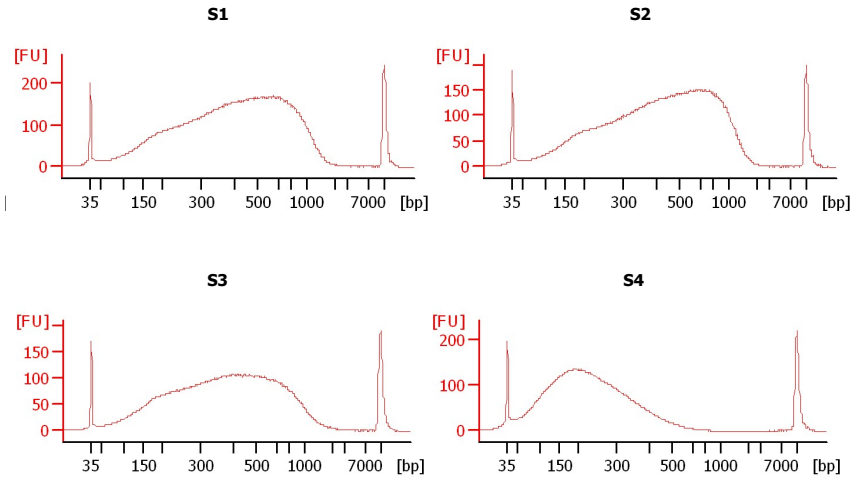


Figure 16. Fragment length distributions after Bioruptor fragmentation. The samples labelled "L" are prepared starting from LCL DNA, the ones labelled "S" from saliva. The sonication cycles number is indicated by the numerical tag in the samples name: 1=5 cycles, 2=10 cycles, 3=20 cycles, 4=30 cycles. Distribution peaks are: L1=1212bp, L2=1307bp, L3=408bp, L4=556bp, S1=576bp, S2=595bp, S3=392bp, S4=197bp.

While saliva samples showed overall shorter fragment lengths (197-595bp for saliva samples, 408-1307bp for LCL samples), the fragmentation appeared to be inconsistent, with the distribution peaks slightly uncorrelated to the cycle number in both saliva and LCL samples (S2 and L2 fragments are longer than the corresponding S1, and L1). The distributions of fragment lengths result broader in LCLs at

high cycle numbers, while saliva samples showed a reverse scenario in which the higher cycle number fragmentation produced narrower distribution.

A second test was performed using the same settings as the first, with the exception of the intensity, which was set at 320W ("HIGH" setting). The samples of this second test were only saliva DNA, being this the type of sample that will be used for library preparation and target enrichment testing, but repeating the same cycle number in two samples instead of one (Figure 17). Again, the results showed incongruence between cycle number and fragment length distributions. The distribution of fragment lengths, that show a longer tail towards shorter measures, become more gaussian-like after 20 cycles of sonication.

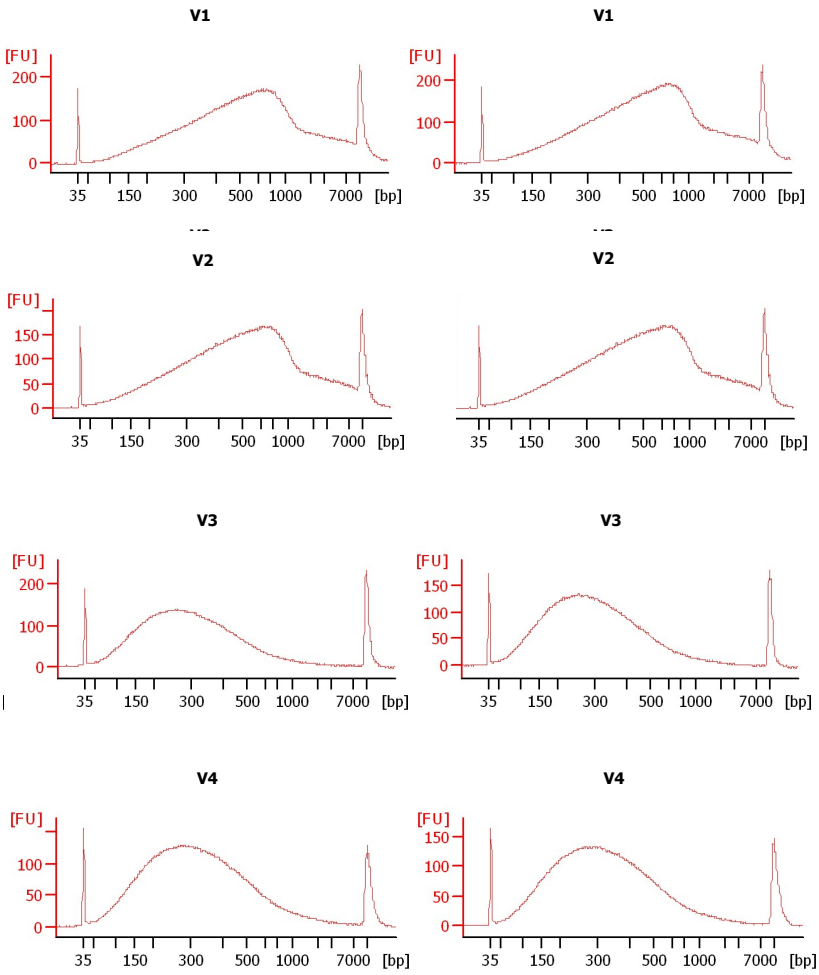
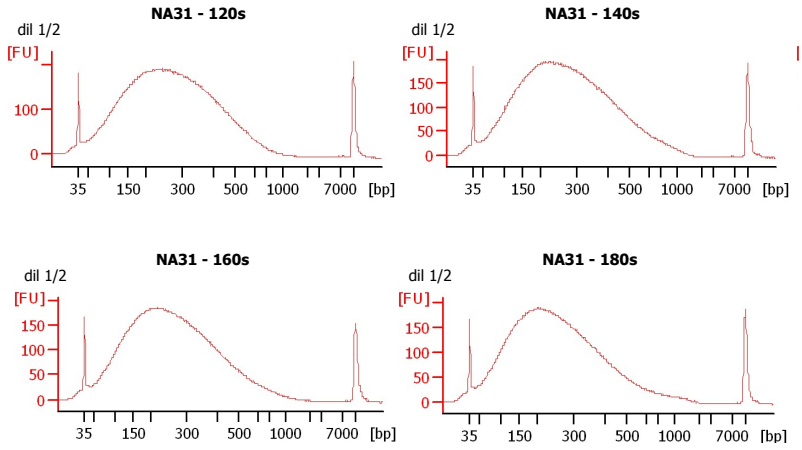


Figure 17. Fragment length distributions after Bioruptor fragmentation. The sonication cycles number is indicated by the numerical tag in the samples name: 1=5 cycles, 2=10 cycles, 3=20 cycles, 4=30 cycles. All fragmentation were repeated twice per sample. Distribution peaks are: V1=626bp and 641bp, V2=670bp and 649bp, V3=258bp and 258bp, V4=286bp and 273bp.

The variability in the Bioruptor fragmentation results drove us to opt for a Covaris system fragmentation instead. The testing was performed again on two samples sets: LCL DNA and human saliva DNA samples, testing 120, 140, 160, 180 and 200 seconds of exposure. The measured fragment length distributions showed very little difference in LCL samples, and almost the same fragmentation patterns in all saliva samples (see Figure 18). The saliva samples peaked around 180bp, which was a little shorter than the aimed fragment length, leading to a last fragmentation experiments using lower time of exposure (80s) (Figure 19). The distribution of fragment lengths using 80 seconds of sonication time was wide, but with a strong peaking region at 200-400bp, the optimal range for our experiments.



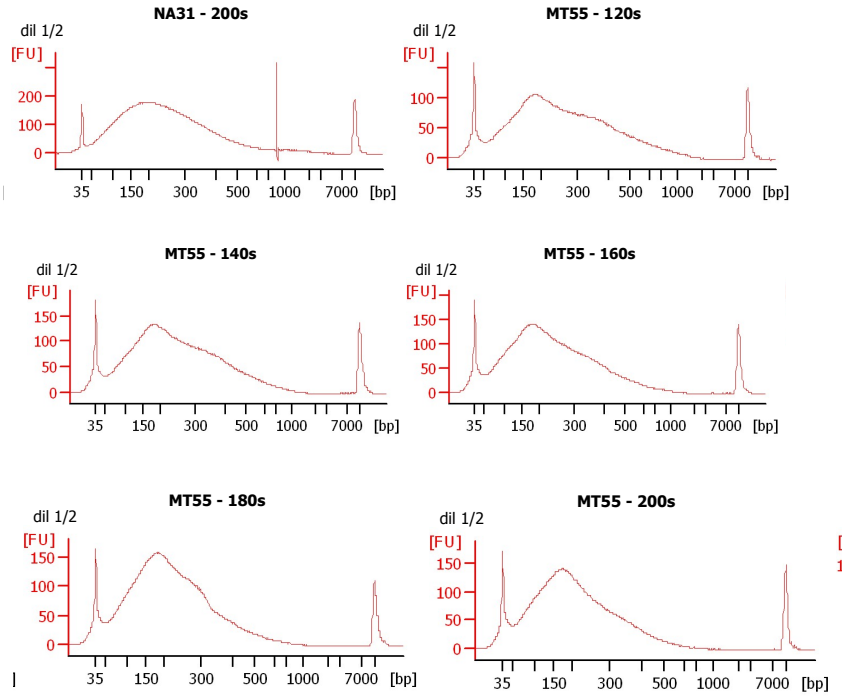


Figure 18. Fragment length distributions after Covaris fragmentation. The samples labelled "NA31" are prepared starting from LCL DNA, the ones labelled "MT55" from saliva. The sonication time in seconds is indicated by the numerical tag in the samples name. Distribution peaks are: NA31-120s=245bp, NA31-140s=226bp, NA31-160s=219bp, NA31-180s=203bp, NA31-200s=208bp, MT55-120s=182bp, MT55-140s=181bp, MT55-160s=184bp, MT55-180s=181bp, MT55-200s=177bp.

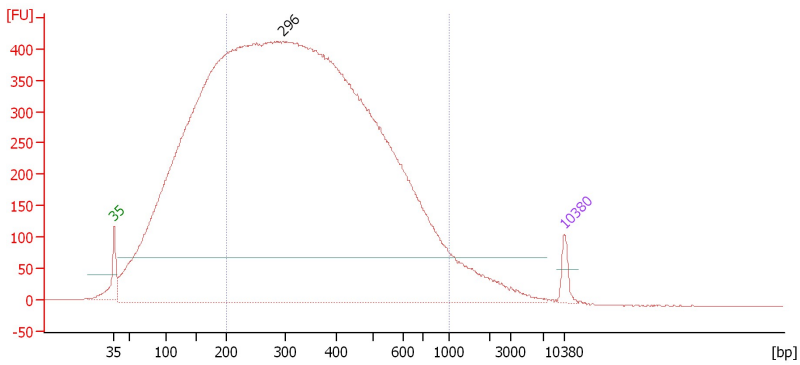


Figure 19. Fragment length distribution after Covaris fragmentation setting the sonication time at 80 seconds. The distribution shows a peak at 296bp.

Libraries were produced starting from 14 human saliva samples. The samples were tested for EBV presence using real time PCR, and showed large variation in virus copy concentration (see section 4.4.2.). The samples were chosen at random in order to simulate the virus presence in a real sampling data set.

In order to control that the library protocol worked correctly, a positive control (PCR amplicon) have been added as additional sample to the sample set. An aliquot of the positive control was taken before the beginning of the library preparation, after the adapter ligation, and after the indexing

PCR. If correctly processed, the amplicon should show an additional ca. 60bp after the adapter ligation (each PE adapter is ca. 30bp), and 60bp more should be added during the indexing PCR. The amplicon length showed the expected pattern, proving the correct performance of the protocol (Figure 20).

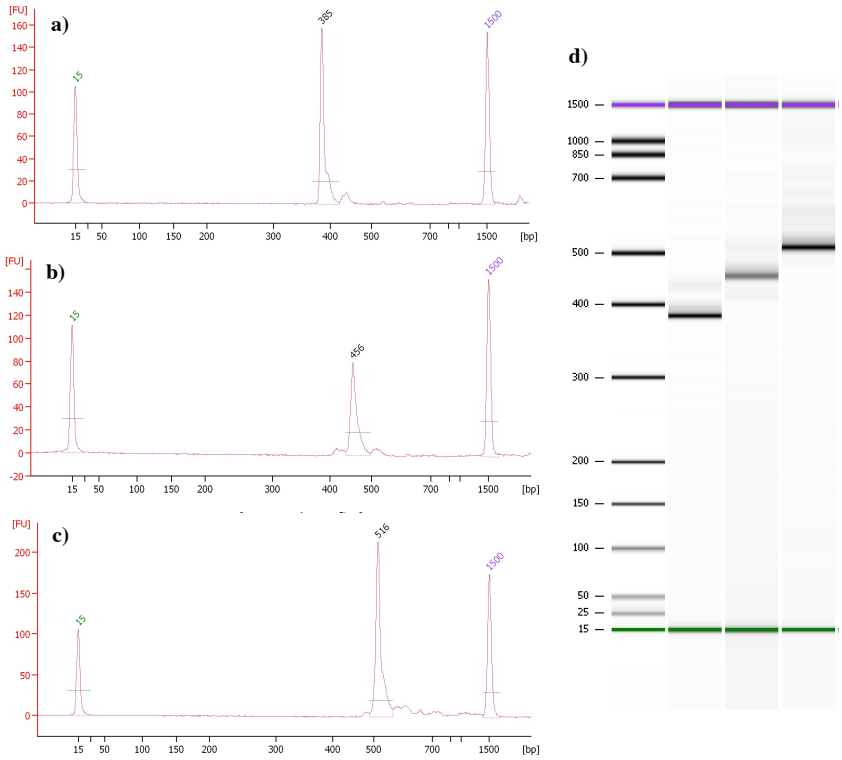


Figure 20. Fragment lengths of the library preparation positive control (amplicon) during processing. a) Amplicon before library preparation beginning. b) Amplicon after adapter

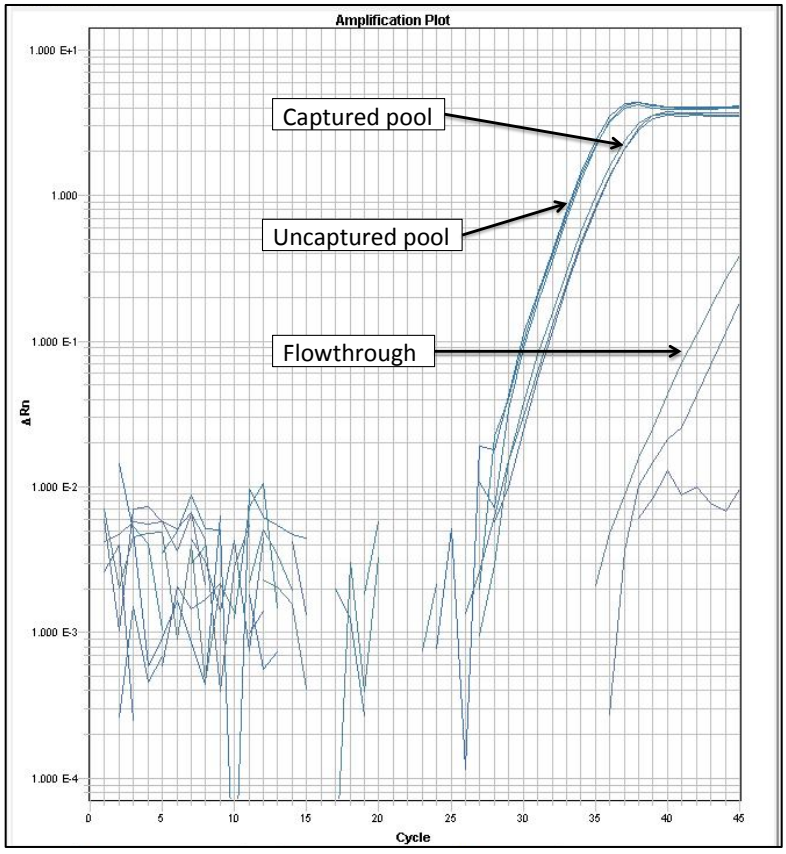
ligation. c) Amplicon after indexing PCR (complete library). d) Virtual electrophoretic gel representation of the 3 samples.

The final libraries were quantified using Thermofisher Qubit Fluorimeter and equimolarly pooled for target enrichment.

4.4.2. Target enrichment

The samples pool undergoes target enrichment, and aliquots of the pool were taken before and after the enriching (before the post-enrichment PCR in order to have comparable samples for virus presence). An additional aliquot was taken of the target enrichment flow through (*i.e.* what remains of the sample after being depleted of the enriched target through the paramagnetic beads). The aliquots were then tested through real time PCR for relative quantification of a region of EBV, and Human's genomes. The results are shown in Figure 21.

Human DNA content



EBV DNA content

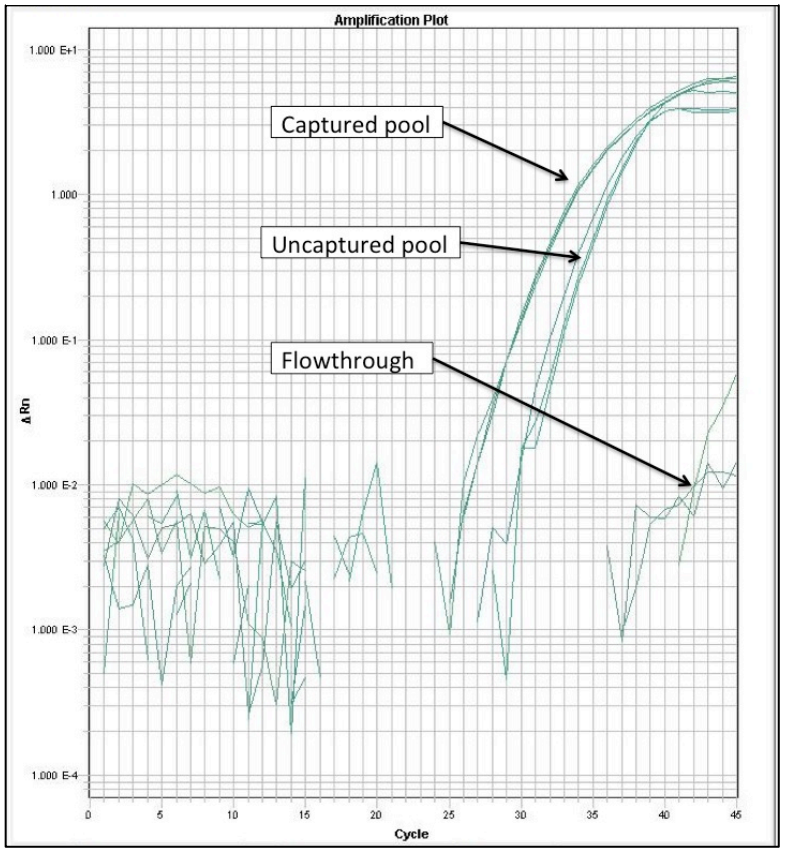


Figure 21. Real time PCR of the target enrichment test samples. This techniques measures over time the amplification of a specific region through the cumulative fluorescence given by probes that are able to attach only to the amplified region itself. The X axis shows the number of PCR cycles, while the Y axis shows the intensity of the fluorescence. The experiments were performed in triplicates, hence the three lines per sample. The earlier the fluorescence intensity increases and reaches a plateau, the higher the amplified region content. In the Human amplicon amplification, it is possible to see how the Human DNA content is lower in the "captured pool" (*i.e.* the enriched sample pool) compared to the uncaptured one (*i.e.* the

sample pool before target enrichment. In the EBV amplicon amplification, the results are expectedly inverse, with a higher amount of target amplicon compared in the captured pool compared to the uncaptured or the flow through.

The content of Human DNA appears to be reduced in the sample pool after enrichment, while the viral content increased. This indicates that the target enrichment was successful, with an high increase of EBV presence after the enrichment.

The enriched sample pool was sequenced, and the obtained reads mapped against EBV, HHV-6A, HHV-6B and Human genomes. The reads span in an approximately uniform fashion along the genome of the targeted viruses, unrelated to GC content, which indicates the absence of strong bias towards specific regions (see Figure 22).

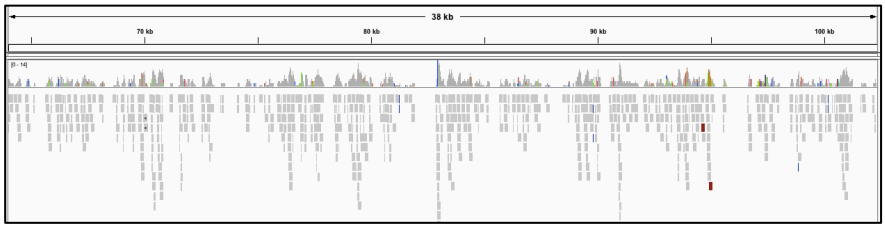


Figure 22. Representation of the mapping reads positions of one of the samples after enrichment in a random region of the EBV genome. The image was produced with the Integrative Genome Viewer(Robinson et al., 2011).

The sequencing results, summarized in Table 8, show little variance in the number of reads outputted per individual, indication of a correct equimolar pooling of the samples.

a)

Sample	Total Reads	Representation (%)	On target(%)	Unmapped(%)	Duplicates(%)	Human(%)
DH2S	187918	8,5	0,0106	99,99	0,0	89,5
DI3S	70824	3,2	0,0226	99,98	0,0	66,5
JR3S	139339	6,3	0,1069	99,89	0,004	78,9
MB1S	94798	4,3	0,0264	99,97	0,0	86,6
MOR009	111032	5,0	0,0108	99,99	0,0	88,9
MOR015	136092	6,2	0,2697	99,73	0,0	45,6
MOR025	204176	9,2	0,0039	100,00	0,0	84,9
MOR030	228236	10,3	1,1422	98,86	0,023	68,6
MOR043	223250	10,1	0,0157	99,98	0,0	82,0
MOR066	243818	11,0	0,0701	99,93	0,0	79,4
MOR122	148208	6,7	0,0088	99,99	0,0	84,5
MOR198	197492	8,9	0,0051	99,99	0,0	89,9
MT8S	70220	3,2	0,0299	99,97	0,0	82,5
NAD016	155358	7,0	0,0071	99,99	0,0	88,7

b)

Sample	EBV (%)	HHV-6A (%)	HHV-6B (%)
DH2S	0,0011	0,0032	0,0064
DI3S	0,0085	0,0056	0,0085
JR3S	0,1048	0,0007	0,0014
MB1S	0,0011	0,0074	0,0179
MOR009	0,0000	0,0045	0,0063
MOR015	0,2535	0,0059	0,0103
MOR025	0,0000	0,0010	0,0029
MOR030	1,1335	0,0044	0,0044
MOR043	0,0004	0,0076	0,0076
MOR066	0,0648	0,0016	0,0037
MOR122	0,0000	0,0040	0,0047
MOR198	0,0015	0,0015	0,0020
MT8S	0,0128	0,0071	0,0100
NAD016	0,0000	0,0039	0,0032
Total mapped	0,15	0,004	0,006

c)

Sample	Copies*100ng	Est.copies*cell	pre-enrichment (%)	post-enrichment (%)	Enrichment (X)
DH2S	106,3	0,00744	1,87E-05	0,0011	57
DI3S	171,4	0,0	3,02E-05	0,0085	281
JR3S	1125,1	0,07875	1,98E-04	0,1048	529
MB1S	2,9	0,00020	5,05E-07	0,0011	2088
MOR009	7,3	0,00051	1,29E-06	0,0000	0
MOR015	3387,8	0,2	5,96E-04	0,2535	425
MOR025	0,8	0,0	1,40E-07	0,0000	0
MOR030	17582,8	1,2	3,09E-03	1,1335	366
MOR043	0,3	0,0	5,58E-08	0,0004	8031
MOR066	435,5	0,0	7,67E-05	0,0648	845
MOR122	0,5	0,0	8,41E-08	0,0000	0
MOR198	0,8	0,0	1,46E-07	0,0015	10390
MT8S	98,7	0,00691	1,74E-05	0,0128	738
NAD016	1,0	0,0000733	1,84E-07	0,0000	0

Table 8. Target enrichment results. a) Percentage of reads per individuals over the total reads are shown as "Representation". The sum of the mapped reads of EBV, HHV-6A and HHV-6B (*i.e*. on target) are calculated as percentages of the total reads per individual. b) Percentages of mapping reads per individual to the three enriched viruses, and total reads mapping per virus. c) EBV target enrichment results. The virus copies for 100ng of sample DNA, for cell, and the pre-enriched EBV DNA amount were calculated using quantitative real time PCR. The post-enrichment viral DNA amounts were estimated as the mapping reads over the overall reads per individual. Marked in yellow are the individuals showing more than 100 mapping reads.

The reads mapping to one of the targeted virus, or "on target" reads, per individual vary from 0.004-1.14%. When counting these reads, EBV appears to have 25 to 40-fold higer values than HHV-6A and HHV-6B, but most of the reads belong to a single individual. This individual is MOR030, the one showing the vastly higher value of EBV copies in saliva before the

enrichment. When MOR030 is removed from the comparison, the overall EBV mapping reads value lowers to 0.04, a value that is nevertheless still 10-fold higher than the one of the other two viruses. Also, the variation in amount of virus reads per individual has little variation in the two HHV-6 species compared to EBV, sign of a more stable viral copy number in saliva.

The most relevant result of the capture is the enrichment achieved, a value can be estimated only knowing the original samples virus copy number. This would then be compared to the ratio of reads mapping to the virus after the enrichment per each individual over the total individual reads per individual, as a proxy for viral copy number. Almost all individuals showed less than 10 reads mapping against HHV-6A or HHV-6B, which makes the calculation statistically weak, could not be used for such purpose. EBV showed instead five individuals with higher number of mapping reads (marked in yellow in Table 8c), which we consider more solid data to determine the enrichment achieved. We then took advantage of our quantitative real time PCR designed for EBV to

calculate the virus copy number in the samples before the capture, with results that unsurprisingly varies wildly.

While the on target reads, number appears small compared to the unmapped one, the tiny dimension of EBV genome, and the relatively low amount of copies in a biological sample implies that even a <1% viral DNA/total DNA can be a considerable improvement. The overall enrichment ranged in fact between 366-845X.

The PCR duplicates levels were very reduced, a result partly given by the low number of mapping reads for most individuals, but that remains very low also in the individuals showing higher on target read numbers.

The reads that were not mapping against any of the enriched viruses were mapped against Human. The mapping results showed 66-90% Human origin of the sequenced reads, and a Megablast-based taxonomic examination showed that the rest are from a wide range of microbes, an expected result being the samples saliva total DNA.

The overall enrichment was high, with a contained cost due to the in-house library preparation and the multiplexing during the hybridization phase. Nevertheless, these values are still too low to use the method to generate complete sequences, and for a large-scale study, but it is a well-placed foundation from where to build forward. To improve the method the testing of additional techniques is needed, such as human DNA depletion through inverse target enrichment (generate in-house human genomic baits and target them to extract them from the sample to use for viral enrichment), or successive double hybridization is the next step to improve the method.

5. DISCUSSION

5.1. Data generation

Herpesviruses are ubiquitous and infect the almost totality of the adult human population. While generating genetic data for these organisms during their lytic replicating life cycle phase is relatively easy, this becomes a complex challenge for their persistent infection forms. Latent infections by both HHV-4 and HHV-6 are characterized by very low viral loads, estimated for EBV to range between 1-50 copies per million cells(Babcock et al., 1999). To directly sequence DNA samples latently infected with these viruses would produce almost no viral reads compared to the human-mapping ones due to both a low viral load and a small virome size (ca. 160-170 kb in HHV-4 and HHV-6). Hence more specific approaches such as PCR amplifications or target enrichment are required. The development of protocols aimed to generate data from latently-infected samples would increase greatly the potential source of infected individuals to study due to the high infectivity and long-life infection typical of these viruses.

In this work, we explore different approaches to generate data, depending on the target virus form, the sample type and the number of available samples.

5.1.1. Inherited chromosomally integrated HHV-6 target enrichment

The 1000 Genome Project made available a large amount of data derived from shotgun sequencing of thousands of individuals. The data was produced starting from blood and Lymphoblastoid Cell Line (LCL) samples. Blood samples include different cell types potentially infected by latent HHV-6 due to the transmission in the bone marrow before their differentiation(Luppi et al., 1999). LCL samples can also be recipient of HHV-6 latent infection, lymphoblasts being one of the cell types for which this virus has high tropism in all its forms(Luppi et al., 1999; Takahashi et al., 1989). Since latent HHV-6 integrates in very few of the host cells(Morissette & Flamand, 2010b), the probability to culture an infected one in an LCL is low, and moreover these lines end up being oligoclonal or monoclonal. The inherited chromosomally

integrated form of HHV-6 (iciHHV-6) is instead passed through generations in a Mendelian fashion, causing the offspring to have a copy of the virus in each cell. Individuals with this condition would generate LCLs with high HHV-6 viral load. The prevalence of iciHHV-6 in the overall population ranges around 0.2-1%(Leong et al., 2007; Tanaka-Taya et al., 2004), which is consistent with the number of samples we found having reads mapping to HHV-6 within the 1000 Genome Project (approximately 0.5%). Even with a copy of HHV-6 genome per cell, the coverage achieved for HHV-6 from these samples never exceeded 2.5, which is not sufficient to perform solid variability analysis. This led to the decision to perform target enrichment on the biological samples derived from the putative iciHHV-6 carriers.

Large repeats impair the uniform representation of the whole virome because of probe saturation in target enrichment, and because of the short nature of NGS sequencing reads that impairs their reconstruction. The chosen target enrichment design was consequentially built excluding these regions,

leaving us with the whole unique region of HHV-6 virus, which amounts to ~87% of the whole virome.

The target enrichment results were satisfactory, with uniform coverage in all iciHHV-6 sequences and mean values ranging from 77-229X. At the time of production of this data, this was the largest and most complete whole-genome data set produced for iciHHV-6, as well as for overall HHV-6.

While we could infer the presence of iciHHV-6 in the study samples, the exclusion of unlikely alternative scenarios, such as the integration of the virus in few B-cells of the host followed by a clonal expansion of the infected cells upon transformation and LCLs establishment cannot be ruled out without sampling other tissues from the same individual, but can certainly be considered very improbable. While the standard method to detect iciHHV-6 is through the combined analysis of whole blood viral load(Luppi et al., 1993; Ward et al., 2006), FISH and virus-chromosome junction amplification(Arbuckle et al., 2010; Nacheva et al., 2008), we took advantage of the recently described ddPCR-based

method. Published by Sedlack *et al*.(Sedlak et al., 2014) this protocol is efficient, simple, and cost-effective, and allowed us to prove the presence of an integrated HHV-6 copy per cell, thus proving that the infecting form was indeed iciHHV-6.

5.1.2. EBV latency genes amplification

A set of genes were sequenced from a large panel of EBV latently infected individuals, in order to study variability and stratification. While neutral regions would be more fit for this kind of analysis, the presence in the EBV genome of very few introns, and of little non-coding or regulatory RNAs regions that are not pervaded by repetitive elements, led to the choice to amplify whole genes, which would allow us to work with a larger amount of data and also potentially test for evidence of selection.

The sample panel included more than 350 samples from individuals with diverse geographical origin. While the different origins of the samples would allow for geographical stratification analysis, the crucial characteristic of this panel was the type of sample: saliva from healthy individuals.

These individuals did not present lytic infection at the time of collection, and the almost ubiquitous prevalence of EBV hints to the presence of a latent infection. Saliva is the main transmission medium for EBV, and it is where the chronically-infected pharyngeal cells produce constant shedding. Saliva has consequently higher latent EBV viral loads than most of the other body tissues or compartments(Hadinoto et al., 2009), increasing the chance of success of the amplifications.

The targeted regions for amplifications belonged mostly to latency genes. These genes are the most variable in EBV(Palser et al., 2015; Sample et al., 1990; Santpere et al., 2014), allowing for higher resolution analysis because of the sheer number of SNPs. Furthermore, the distribution of these genes covers very different genome positions, which reinforces the results of the stratification analysis in a virus such as EBV, where recombination is pervasive(Palser et al., 2015; Santpere et al., 2014; Walling & Raab-Traub, 1994). To study a single region of the EBV genome could bias this kind of study if the region itself has been recombined between strains, an effect that is limited when analysing more regions

at once. Lastly, EBV latency has been identified as having a substantial role in many human disorders, with the latency gene products being strongly related to the virus ability to induce malignancies(Kang & Kieff, 2015), increasing the interest for the characterization of this set of genes.

The complete design of targeted regions sums up to a total of almost 40 kbp, a magnitude that can be reached by short PCR amplicons only using a large number of reactions. This process would be very laborious and time-consuming, which brought to the decision to use instead long-range PCR coupled with NGS techniques. This allows for a limited number of reactions, which is an important factor when taking into account the large sample size and the relatively large amount of DNA needed for each amplification. Using an in-house protocol to generate sequencing libraries from the amplicons, and multiplexing all the individuals in a single pool, allowed for a time- and cost-effective data generation.

The amplification and sequencing results showed good coverage in the selected 102 individuals, with few genes not represented in random individuals. Within these cases, the

EBNA2 region was the more conspicuous, failing to reach the quality thresholds for 71/102 individuals. This led to its exclusion from the analysis, resulting in a final data set with less than 10% of missing data. This includes all the other latency products in one of the largest, and the most diverse, EBV data set produced to date.

5.1.3. Cost-effective target enrichment protocols

Target enrichment has become a key technique in genomic analysis, being increasingly used in a vast range of fields. Its plasticity and effectiveness allows to sequence target DNA by reducing the regions not of interest. The scalability and sensitivity of this technique makes it one of the main alternatives to more laborious methods such as amplification of large genomic regions through small amplicons.

Target enrichment has been used extensively in virology, where often the human genomic background in a sample is relatively high compared to the virus one, even in a sample derived from an active infection. While this technique was used until recently mainly to target small RNA viruses, it has

increasingly been used to sequence larger ones, within which examples can be found in various herpesviruses such as HCMV(Daniel P. Depledge et al., 2011), HHV-7(Donaldson, Clark, Kidd, Breuer, & Depledge, 2013), VZV(D. P Depledge et al., 2014) and HSV-1(Ebert, Depledge, Breuer, Harman, & Elliott, 2013). With very few exceptions, the enrichments were always performed on samples with high viral load collected from patients affected with different virus-related diseases, usually suffering viral reactivation. Target enrichment attempts on healthy individuals, where the virus is in its latent infection form, are very few. In spite of that, the exploration of this technique to generate latent virus genomes is relevant to study the most common source of biological samples (saliva or blood). Also, to generate in a cost-effective manner the large number of sequences needed to study population variability and stratification, it is important to facilitate collection and improve the possible sample source in terms of individuals. Very often the cohorts studied are limited by the selection of individuals with high viral loads. To be able to target latent viruses in herpesvirus-infected individuals would in principle significantly improve the sample size thanks to the

almost ubiquity of this class of viruses. In addition to this, the possibility to use saliva for virus target enrichment would reduce the effort of sampling, compared to more invasive type of samples such as blood. The generation of data set built from the same sample type is furthermore important for comparison between studies (to date the sample types used to study EBV are many, and often multiple in a single study), because of the comparability of the results, and because of the proven different preferential compartment fro infection from certain variants in an individual(H. . Chen, Lung, Chan, Griffin, & Ng, 1996; Gutierrez et al., 1997; Sacaze, Henry, Icart, & Mariame, 2001; Triantos, Leao, Porter, Scully, & Teo, 1998).

Whole genome analysis are very relevant when studying viruses such as EBV. Extensive diversity has been evaluated for this virus, which is heterogeneous in different regions of the virus, as well as for the multiple strain infection show to be carried by most human hosts(M. . Bell et al., 2008; Bhatia et al., 1996; Edwards, Seillier-Moiseiwitsch, & Raab-traub, 1999;

Gutierrez et al., 2002; Sitki-green, Covington, & Raab-Traub, 2003; Walling, Brown, Etienne, Keitel, & Ling, 2003). Also, as shown in Gutierrez *et al*.(Gutierrez et al., 2000), there is incomplete linkage of genetic variation observed across different regions, suggesting that EBV's heterogeneity is likely higher than what has been estimated.

Target enrichment has been already used to generate sequences both in EBV and HHV-6. Still, the most used technique for HHV-6 whole-genome sequencing had been the generation of long-range overlapping amplicons combined with NGS technology(Joshua Tweedy et al., 2016; Zhang et al., 2017). This technique, based on a 40kbp long range amplicons design, allowed for high quality sequence generation, but with relatively high costs. Also, it has been applied to individuals affected by iciHHV-6, who as such present an abnormally high viral load compared to normal latently-infected individuals. As an alternative, in a study currently available in BioRxiv instead, Greniger *et al*.(Greninger et al., n.d.) applied target enrichment to generate HHV-6 sequences in a cohort composed of 121

individuals. The study produced high-coverage sequences representing 99% of the genome for the B variant and 40-78% for the A variant (the design was built solely on the B variant reference). Nevertheless, Greninger *et al*. focused again the study on iciHHV-6-infected individuals, leaving unresolved the question of the efficacy of the application of this technique in exogenously-infected ones.

Target enrichment has been used on EBV mainly to enrich samples in which the virus was in primary infection or in the reactivated state, two states characterized by high viral loads. Depledge *et al*. enriched EBV in samples showing virus reactivation, producing two partial (>95% of genome covered) high-coverage sequences in 2011(Daniel P. Depledge et al., 2011). In a larger scale, Palser *et al*. enriched EBV from samples with viral loads above 10^6 total copies(Palser et al., 2015). The sequences cover the totality of the unique region of the virus. The overall enrichment achieved in this study averaged 2000-fold, but it required extensive sequencing, with 6-25 samples multiplexed per lane on an Illumina HiSeq 2000. In this same study, a single EBV sequence was

produced starting from the saliva of a healthy individual. It is not specified the number of samples multiplexed with this saliva-derived sample, as well as the sample viral load. Nevertheless, the sequence obtained from it had overall high-quality, and is a positive sign of the feasibility of capturing EBV from healthy individual saliva.

The testing performed in this study resulted in high enrichment for EBV, ranging from 366-845-fold, values that could not be estimated for HHV-6 because of the limited number of mapping reads, and the consequent low statistical robustness. This might imply that exogenous HHV-6 latency is characterized by viral loads that are too low to generate whole genome sequences, at least in a cost-effective way. It is worth pointing out that the sequencing of the target enrichment was achieved using an Illumina MiSeq Micro Kit V2, a kit that outputs a very limited maximum of 4 million reads, and is thus often used for testing, limiting the estimate of the actual enrichment on the final data. The adapted protocol was successful in different aspects of the capture.

The final representation of the multiplexed individuals expressed in number of reads per individual was satisfactory, with a minimum and maximum at 3.2 and 11% of the total reads, respectively. The proportion of PCR duplicates was also very low, never exceeding 1%. While this value might be underestimated because of the limited number of reads, it does not vary significantly in individuals with higher number of mapping reads. Lastly, the coverage was uniform along the EBV genome, again confirming the correct probe design.

The EBV viral loads for the saliva of the tested individuals varied greatly. Even using relatively large amounts of DNA, the range of total viral copies per reaction was significantly lower than the 10^6 lower thresholds set by Palser *et al*. in their capture-based study, with reliable measurable enrichment for samples having less than 1000 total copies.

Overall, the target capture we tested has shown that this technique has the sensitivity to enrich viruses in very low copy number, at least for EBV, and that the data obtained are suitable to generate complete sequences with relative coverage uniformity and low PCR duplicates. Nevertheless,

even taking into account sequencing on high-output platforms, the achieved enrichment is still not sufficient to make this protocol cost-effective for large-scale studies.

The optimization of the protocol will continue following different strategies to improve its efficiency. In order to improve the overall enrichment, the main modification to the protocol to test is to repeat the capture sequentially on the samples. Using this strategy, the samples undergo target enrichment, and the enrichment pools undergo a second round of enrichment in order to decrease the human DNA content. Another approach would be to perform target enrichment on the sample pools using baits designed to capture specifically human genome. This would deplete the sample pool of human DNA, increasing the virus representation. Lastly, following the example of Palser *et al*., a viral load quantification of the samples should be used to pool samples with similar viral copy number in order to limit the unbalanced representation of the single strains in the sequencing results.

5.2. Data analysis

5.2.1. HHV-6 variability in its exogenous and congenital forms

iciHHV-6 is characterized by the presence of at least a copy of the virus integrated in the genome of each cell of the host. The integrated viruses are intact and complete, and evidence for expression/reactivation substantially hint to their functionality(Joshua Tweedy et al., 2016). While the presence of this unusually high viral load is known to have problematic outcomes in solid organ or bone marrow transplants, the link between this type of infection and other putatively associated diseases is still unclear. A fundamental step to shed light on these links is to characterize the genetics of iciHHV-6, as well as its difference with exogenous HHV-6.

Since the hypothesis was formulated trying to explain the presence of such high HHV-6 viral loads in random individuals without a lytic infection, much has been studied of what we now know to be the congenital form of the virus. iciHHV-6 was until recently identified in patients using a combination of Fluorescent In Situ Hybridization (FISH) and

PCR amplification(Arbuckle et al., 2010; Nacheva et al., 2008). While the viral load in whole blood was used as a sign of iciHHV-6 infection in healthy individuals, where values above 5.5 $\log_{10}$ copies/ml(Caserta et al., 2010; Clark et al., 2006; Deback et al., 2009) would be an indicator of its presence, this was not considered a conclusive diagnostic result. The selected individual would then be tested by FISH, a technique that requires live cells to be performed, and long-range PCR amplification of the junction between virus and chromosome. This only required to have DNA extracted from the cells rather than the cells themselves but was complicated by the different primers sets and conditions to be used with the different chromosomes. Different additional tests were designed, but were not conclusive, and/or needed uncommon samples such as nails or hair(Ward et al., 2006). The different invasive (blood) and uncommon (nails, hair) samples needed for these tests, and their technical challenge, hindered large-scale iciHHV-6 detection, even though some studies had been able to overcome the time and money limitations(C. Hall et al., 2004; Leong et al., 2007). This changed with the introduction of ddPCR in iciHHV-6 studies, demonstrating

how this precise and easy technique would allow for congenital integration detection in a rapid and cost-effective fashion that required minimal DNA quantity(Hill et al., 2016; Sedlak et al., 2014), although it has the drawback that it does not provide information on the specific insertion point of the virus.

We took advantage of ddPCR to confirm with high level of confidence the hypothesized congenital integration in the selected samples from the 1000 Genome Project. The result showed that all cases we selected were very likely positive for iciHHV-6. The hypothesis we developed in this study was that it is possible to detect iciHHV-6 from low coverage shotgun sequencing of tissues such as whole blood or LCLs. Screening the phase 3 sequences of the 1000 Genome Project with a selective threshold set at median HHV-6 coverage >=2, we identified 11 putatively infected individuals. Only 9 of those presented available biological samples in the Coriell's Institute cell-lines repository, which we tested through ddPCR confirming the correct assignation of our test. We do not argue that our test is diagnostic for iciHHV-6, but it

could be used to identify putative candidates in large number of sequences, reducing the number of ddPCR tests for confirmations.

At the moment of submission, our work presented the largest comparative whole-genome analysis of both iciHHV-6 and overall HHV-6. Being two different species(D. Ablashi et al., 2014), the variability of the two variants of HHV-6 has been evaluated separately and compared. The diversity, consistently with what described using mainly 4 genes, was found to be significantly lower in the B variant(Joshua Tweedy et al., 2016), confirming the divergence with the A variant. iciHHV-6 diversity was lower than the overall HHV-6 one, a result possibly given by the divergence between the two forms, which are indistinctly included in the overall data set. To confirm this hypothesis, a larger exogenous HHV-6 data set would be needed, again calling for methods to generate more sequences. When the genome features (exons, introns, UnTranslated Regions (UTRs), and others) were compared, the ratio of variability remained constant

between the two variants, with introns showing the highest diversity by far. This result could be explained by the evolutionary indifferent nature of this genomic feature, and the consequent lack of selective pressure acting upon it.

An important observation that arose from the variability analysis is the identification of the HHV-6B reference sequence Z29 as "uncommon". An organism reference sequence is expected to be genetically within the ‘boundaries’ of the genetic polymorphisms present in that species, a scenario that was not found in our data set. The distribution of polymorphisms among individuals showed a disproportionate number of exclusive variants in Z29 compared to the other HHV-6B sequences, including the only non-sense mutation found. This mutation, classified by reference-mapping as a stop-loss, was present in all HHV-6B data sets, implying the possibility that it could instead represent a stop-gain mutation acquired by Z29. Principal component analysis and phylogeny identify Z29 as the most genetically divergent sequence, behaving in phylogenetic trees almost as an out-

group. Z29 is the present HHV-6B reference sequence solely because it was the fist complete sequenced strain of this species, raising the question of weather it is the best candidate for the reference role. Again, the limited number of complete HHV-6B sequences hinders a clear answer to this question.

Geographical stratification can be strong enough in HHV-6, to confound any results of genetic analysis trying to link viral polymorphism to diseases, if not properly taken into account. The main pathologies associated to HHV-6 show strong geographical patterns(Ascherio & Munger, 2010; Huh, 2012), as the prevalence of the virus in different regions(Bhattarakosol, Pancharoen, Mekmullica, & Bhattarakosol, 2001; Linhares, Eizuru, Tateno, & Minamishima, 1991; L. Nielsen & Vestergaard, 1996; Politou et al., 2014; Tolfvenstam et al., 2000; Wu, Mu, & Wang, 1997), hinting to the possibility of region-specific or population-specific variants causative to the diseases. To understand the potential geographical population structure of

HHV-6 would allow to create correct study designs to tackle these two-viral species, and could help to pinpoint specific variants or genomic regions putatively associated with certain diseases.

While the data set of this study was limited in sequence numbers and geographical origins, the combination of PCA, phylogenetic and recombination analysis showed signs of stratification, with an Asian cluster separated from the African individuals in HHV-6A, a scenario resembling the sister-strain EBV(Palser et al., 2015; Santpere et al., 2014). In HHV-6B the results are less straightforward to interpret in terms of population structure, even though signs of the presence of a principal African and European components were found.

The identification of instances of evolutionary adaptation has been a main objective of evolutionary biology since the introduction of the concept of natural selection. At a molecular scale, it is expected that adaptive natural selection would select amino acidic changes that induce a change in the translated protein that would produce a phenotype change.

This implies that non-synonymous positions would be under much stronger selective pressure than synonymous sites. This concept is at the base of one of the most widely used measurements of protein accelerated or slowed evolution: dN/dS. This measure, based on the early work of Kimura in 1977(Kimura, 1977) and the one of Yang & Bielawski 2000(Yang & Bielawski, 2000), is a test based on homologous protein coding sequences. The amount of non-synonymous changes is compared to the synonymous ones in a ratio that would stay around 1 in the complete absence of natural selection. Higher measures of dN/dS, or ω, would imply the action of positive selection, while lower ones the preservative action of purifying selection. The intuitive interpretation of this measure is supported by the theoretical work on the relationship between the ω statistics and the underlying selective pressure in a Wright-Fisher model(R. Nielsen & Yang, 2003). While the concept of dN/dS is relatively simple, the methods developed to estimate it are various and less straightforward. Within those, the most used take advantage of maximum likelihood models for the

estimation of dN/dS in order to estimate the most likely ratio for the whole population starting from a limited number of samples.

Rates of protein accelerated or decelerated evolution were evaluated in our data set, pinpointing a set of genes showing significant signs of positive or negative selection. While part of these genes was described as showing non-neutral molecular evolution rates, we present a more comprehensive list of genes that might be determinant in their stability for the virus survival, latency establishment and replication (conserved genes), or are needed to for rapid variation (positively selected genes). The role of most of these genes in the biology of this virus are still poorly known, and a characterization of their function is needed to understand the selective pressure to which they had been subject during HHV-6 evolution.

Different genes showing the action of positive selection in the B variant had been detected even in the comparison of only the two first sequence published for HHV-6A and -B(Dominguez et al., 1999). These genes, U90 and U100,

have very different role in the virus biology. While the former is an Immediate-Early transactivator, putatively regulating RNA replication, transcription and modification during the first phase of HHV-6 active infection(Dominguez et al., 1999), *U100* is related to HHV-6 entry in the cell and tropism(Pfeiffer et al., 1993; Pfeiffer & Thomson, 1995).

The A variant showed higher variability than the B one, giving higher natural selection footprints detection power due to the sheer number of polymorphisms. In this variants different genes showed high likelihood for positive selection, covering very diverse viral functions such as minimizing immune recognition (U24)(Sullivan & Coscoy, 2008, 2010), virion reconstruction (U11)(Mahmoud et al., 2016) and cell death signal modulation (U95)(Yeo, Isegawa, Chow, & Irol, 2008). The genes showing negative selection constraints were fewer, with the more prominent signal shown by U48, a gene known to be conserved across the whole *Herpesviridae* family.

One of the most constrained genes in iciHHV-6A was expectedly *U48*, the gene encoding for the glycoprotein H of

the gL-gH-gQ complex, and that is known to be conserved across the *Herpesviridae* family.

5.2.2. EBV geographical stratification

As for HHV-6, the patterns of population structure in EBV are an important factor to characterize the genetic architecture of this virus, and could be a confounding factor in disease-association studies based on genetics. In order to evaluate the presence of stratification, we analysed the divergence between EBV sequences derived from healthy individuals in the most comprehensive geographical data set to date. The study is based on a set of cryptic genes related to the asymptomatic state of the virus: the latency genes.

The choice of the target genes selected to be sequenced was based on two key points: 1- Genes with higher diversity and variability would allow a better resolution in determining the distance between sequences. Latency genes are known to be the most variable in EBV, as shown in the studies by Sample

et al.(Sample et al., 1990), Santpere *et al*.(Santpere et al., 2014), and Palser *et al*.(Palser et al., 2015). 2- As it has been shown for many of its sister family members, EBV has been affected by pervasive recombination in its evolutionary history(Palser et al., 2015; Santpere et al., 2014; Walling & Raab-Traub, 1994). This could be a determining confounding factor when analysing comparative sequence divergence in cases where the analysed region is in fact a recombined one, generating a scenario similar to the one of admixed populations. Latency genes are scattered along the whole genome of EBV, limiting this confounding effect because of the simultaneous analysis of different genomic location.

Latency is established and maintained in EBV through a restricted number of genes, which have a cascade effect that modifies substantially the host cell biology. It is then unsurprising that many of the malignancies related to this virus appear to be related to modifications of these genes(Kang & Kieff, 2015), improving the value of new sequences for the scientific community to better characterize their effects on the host cells.

The panel we used to perform the amplification and analysis consisted of healthy individuals saliva samples. While for privacy reasons there is a lack on metadata on these samples, it is known that they are derived from adult individuals. Since these individuals are healthy and adult, and do not present abnormal EBV viral load values, we can assume that the life-cycle stage in which the virus was at the moment of sampling was latency.

The samples were collected from individuals belonging to populations from all continents (Result section 4.3., Table 6), even though the representation of the different location after quality filters was not uniform, with a final lower sample size for Asia. Nevertheless, also after quality filtering the data set remains the most diverse saliva-derived geographical panel on which EBV variability has been studied.

The first analysis we performed on the produced data was a phylogenies based on the whole data set, collapsing the single genes for each individual in a single sequence. Being

homologous recombination a common event in the evolutionary history of this virus(Matteo et al., 2016; Palser et al., 2015; Walling & Raab-Traub, 1994), EBV strains often present regions of their genome acquired from different strains, which includes different geographical origins ones. This leads to a sort of admixture in terms of EBV genomes composed by stretches of diverse geographical origin(Matteo et al., 2016). While in our analysis the number of genes studied is high, their differences in length and variability might result in a stronger influence on the tree from a small subset of the genes, which might fall in recombinant stretches of the genomes, confounding the final result. In order to limit this effect, a consensus tree was built starting from the single phylogenies of each gene. In this analysis, each gene tree has the same impact on the final consensus supertree, limiting the effect of possible components of different origin in the studied virus genomes.

The results do not differ greatly between the two type of phylogenies built, with the supertree showing in a clearer way

the same relationships between individuals showed in the collapsed tree. The African samples were collected in Congo and South Africa. While Congolese strains cluster solidly together, South African resulted more scattered along the trees. South Africa has a relatively recent colonial history, and still today is a country with a population composed of individuals with many different origin. This particular demography makes this result expected. Interestingly, most of the South African individuals showed little divergence with Argentinian strains.

The strongest signal of population structure found in all stratification studies on EBV genomes is the Africa-Asia separation. In our data set the very limited Asian sequences, all originated from China, hampers the exploration of this separation, which would require the inclusion of the same set of genes here taken in consideration from Palser *et al*.'s and Matteo *et al*.'s studies(Matteo et al., 2016; Palser et al., 2015), which will be the next step in this project.

The South American strains cluster in a similar way to the Africans, with Bolivian ones clustering solidly together, while

Argentinian ones divided in small clusters along the trees, often associated to South African strains.

Lastly, Middle Eastern sequence form different large clusters align the tree. Being the population with the highest number of sequences in our study, random formation of cluster is expected, but their strong domination in these clusters lowers the likelihood of such scenario.

Overall, the phylogenies show the presence of geographical structure, adding to the known population of this virus a Middle Eastern one. The comparison of these results with the published data on EBV stratification is hindered by the different origins of the analysed strains. To put in context the data produced in this study, a comparative analysis comprising all EBV strains is due. Not only this would enlarge the diversity of the data set in terms of populations, but will allow to understand how fine the resolution of stratification analysis can get using a set of genes compared to whole genomes.

The data produced can and will be used to explore other aspects of the variability of EBV, from specific variations in

genes that drives the differentiation of the virus, such as LMP-1 and EBNAs, to the action of selection in shaping its current genetic framework.

5.2.3. EBV viral load stability in Lymphoblastoid Cell Lines

In the Mandage, Telford *et al*.(Mandage et al., 2017) study, the relationship between EBV viral load and host genetic architecture was explored using Genome-Wide Association Analysis (GWAS). The work took advantage of the thousands of published human sequences from the 1000 Genome Project, which are derived from LCLs. These are immortalized (*i.e.* continuously growing and replicating) using EBV, resulting in the co-sequencing of the viral and host genome. While EBV viral load values after immortalization differ substantially between individuals, it is not known whether it is a trait that is stable in time. This stability is an assumption basic to the study of Mandage, Telford *et al*., which had to be proven. Testing the phenotype stability of EBV viral load in LCLs has been my main contribution to this work. The second task I performed in relation to the Mandage, Telford *et al*.

study was the validation of the presented *in silico* viral load estimation method.

The analysis took advantage of the 1000 Genome Project data, which where scanned for EBV-mapping reads. These were used to calculate the coverage for EBV for each sample, and compared to the coverage for the human genome in the same sample in order to estimate the viral load. This method, referred to as *in silico* viral load estimation, had to be tested by comparing the results with experimental measurements in order to evaluate its precision.

The DNA of a set of 13 LCLs samples generated from individuals who participated in the 1000 Genome Project was tested for relative EBV viral load by real time PCR means. The results were compared to the estimation made using Mandage, Telford *et al*.'s *in silico* method, and the correlation between the two sets was calculated. The correlation was high (r2=0.88, p-value=0.00007), proving the validity of the *in-silico* estimation. The possibility to estimate viral loads using

published shotgun sequencing data opens the door to high sample size studies by reducing drastically the costs and time needed for a large number of wet-lab quantification techniques. This is especially relevant in studies such as GWAS ones, where the sample size is the main factor driving the robustness of the results(Hong & Park, 2012).

The stability of EBV viral load in LCLs was tested on 7 LCL derived from 1000 Genome Project's individuals. The immortalized cell lines were cultured, and every 3-4 days part of each culture was removed and its DNA extracted. This resulted in a series of DNA samples representing the cultures at different time points. These samples were tested for relative EBV viral load using real time quantitative PCR and the results compared between cell lines. In order to evaluate the proportion of the variance of the viral load estimate between cell lines compared to the one between time points of a same cell line, two-way factorial ANOVA was chosen. The results showed clearly a significant stronger effect of the different cell lines to the variance, thus proving the stability of

this phenotype, and supporting the subsequent GWAS analysis performed in this study.

Overall, the work presented in this thesis explores different approaches to study variability in *Herpesviridae*. The analyses performed support and extend previous studies results, showing focal aspects of the divergence between strains, and between different viral forms in the case of HHV-6. From these analyses a main limitation arise, the capability to produce viral sequences to generate coherent data sets to allow comparisons between strains and sample types. Because of this, we explore techniques of data production on different samples in a ever more needed cost-effective framework, reaching encouraging results to follow the optimization of protocols aimed to large-scale studies. While yet not conclusive in different aspects, this work sets the bases for more in-depth analyses of HHV-4 and HHV-6 genetics, geographical stratification and evolution.

[Pàgina en blanc]

Bibliography

Ablashi, D., Agut, H., Alvarez-Lafuente, R., Clark, D. a., Dewhurst, S., DiLuca, D., … Yoshikawa, T. (2014). Classification of HHV-6A and HHV-6B as distinct viruses. *Archives of Virology*, *159*(5), 863–870. http://doi.org/10.1007/s00705-013-1902-5

Ablashi, D. V., Balachandran, N., Josephs, S. F., Hung, C. L., Krueger, G. R. F., Kramarsky, B., … Gallo, R. C. (1991). Genomic Polymorphism , Growth Properties , and Immunologic Variations in Human Herpesvirus-6 Isolates. *Virology*, *184*, 545–552.

Adams, A. (1987). Replication of latent Epstein-Barr virus genomes in Raji cells. *Journal of Virology*, *61*, 1743–1746.

Alotaibi, S., Kennedy, J., Tellier, R., Stephens, D., & Banwell, B. (2004). Epstein-Barr virus in pediatric multiple sclerosis. *JAMA*, *291*, 1875–1879.

Arbuckle, J. H., Medveczky, M. M., Luka, J., Hadley, S. H., Luegmayr, A., & Ablashi, D. (2010). The latent human herpesvirus-6A genome speci fi cally integrates in telomeres of human chromosomes in vivo and in vitro. http://doi.org/10.1073/pnas.0913586107

Ascherio, A., & Munger, K. L. (2010). Epstein–Barr Virus Infection and Multiple Sclerosis: A Review. *Journal of Neuroimmune Pharmacology*, *5*(3), 271–277. http://doi.org/10.1007/s11481-010-9201-3

Aubin, J. T., Collandre, H., Candotti, D., Ingrand, D., Rouzioux, C., Burgard, M., … Agut, H. (1991). Several groups among human herpesvirus 6 strains can be distinguished by Southern blotting and polymerase chain reaction. *Journal of Clinical Microbiology*, *29*(2), 367–72. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=269769&tool=pmcentrez&rendertype=abstract

Babcock, B. G. J., Decker, L. L., Freeman, R. B., & Thorley-lawson, D. A. (1999). Epstein-Barr Virus – infected Resting Memory B Cells , not Blood of Immunosuppressed Patients, *190*(4).

Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Ferrell, P. J., Gibson, T. J., … Barrell, B. G. (1984). DNA

sequence and expression of the B95-98 Epstein-Barr virus genome. *Nature*, *310*, 207–211.

Ballestas, M. ., & Kaye, K. . (2011). The latency-associated nuclear antigen, a multifunctional protein central to Kaposi's sarcoma-associated herpesvirus latency. *Future Microbiology*, *6*(12), 1399–1413.

Banwell, B., Krupp, L., Kennedy, J., Tellier, R., Tenembaum, S., Ness, J., … Dilenge, M. (2007). Clinical features and viral serologies in children with multiple sclerosis : a multinational observational study. *Lancet Neurology*, *6*, 773–781. http://doi.org/10.1016/S1474-4422(07)70196-5

Barber, D. ., Wherry, E. ., Masopust, D., Zhu, B., Allison, J. ., Sharpe, A. ., … Ahmed, R. (2006). Restoring function in exhausted CD8 T cells during chronic viral infection. *Nature*, *439*, 682–687.

Bell, A. J., Gallagher, A., Mottram, T., Lake, A., Kane, E. V., Lightfoot, T., … Jarrett, R. F. (2014). Germ-Line Transmitted, Chromosomally Integrated HHV-6 and Classical Hodgkin Lymphoma. *PLoS ONE*, *9*(11), e112642. http://doi.org/10.1371/journal.pone.0112642

Bell, M. ., Brennan, R., Miles, J. ., Moss, D. ., Burrows, J. ., & Burrows, S. . (2008). Widespread sequence variation in Epstein-Barr virus nuclear antigen 1 influences the antiviral T cell response. *Journal of Infectious Diseases*, *197*(11), 1594–1597.

Berti, R., Brennan, M. B., Soldan, S. S., Ohayon, J. M., Casareto, L., McFarland, H. F., & Jacobson, S. (2002). Increased detection of serum HHV-6 DNA sequences during multiple sclerosis (MS) exacerbations and correlation with parameters of MS disease progression. *Journal of Neurovirology*, *8*(3), 250–256. http://doi.org/10.1080/13550280290049615-1

Bhatia, K., Raj, A., Gutierrez, M. ., Judde, J. ., Spangler, G., Venkatesh, H., & Magrath, I. . (1996). Variation in the sequence of Epstein Barr virus nuclear antigen 1 in normal peripheral blood lymphocytes and in Burkitt's lymphomas. *Oncogene*, *13*(1), 177–181.

Bhattarakosol, P., Pancharoen, C., Mekmullica, J., & Bhattarakosol, P. (2001). Seroprevalence of anti-human herpes virus-6 IgG antibody in children of Bangkok, Thailand. *The Southeast Asian Journal of Tropical*

Medicine and Public Health, *32*(1), 143–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11485077

Black, J. B., Sanderlin, K. C., Goldsmith, C. S., Gary, H. E., Lopez, C., & Pellet, P. E. (1989). Growth properties of human herpesvirus-6 strain Z29. *Journal of Virological Methods*, *26*, 133–145.

Borza, C. ., & Hutt-fletcher, L. M. (2002). Alternate replication in B cells and epithelial cells switches tropism of Epstein-Barr virus. *Nature Medicine*, *8*, 594–599.

Brady, G., MacArthur, G. ., & Farrell, P. . (2007). Epstein-Barr virus and Burkitt lymphoma. *Journal of Clinical Pathology*, *60*, 1397–1402.

Braun, D. K., Dominguez, G., & Pellett, P. E. (1997). Human herpesvirus 6. *Clinical Microbiology Reviews*, *10*(3), 521–67. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=172933&tool=pmcentrez&rendertype=abstract

Cagliani, R., Guerini, F. R., Fumagalli, M., Riva, S., Agliardi, C., Galimberti, D., … Sironi, M. (2012). A Trans -Specific Polymorphism in ZC3HAV1 Is Maintained by Long-Standing Balancing Selection and May Confer Susceptibility to Multiple Sclerosis. *Molecular Biology and Evolution*, *29*(6), 1599–1613. http://doi.org/10.1093/molbev/mss002

Caroe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. S., Samaniego, J. A., … Gilbert, M. T. (2017). Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution*. http://doi.org/10.1111/ijlh.12426

Caruso, A., Rotola, A., Comar, M., Favilli, F., Galvan, M., Tosetti, M., … Di Luca, D. (2002). HHV-6 infects human aortic and heart microvascular endothelial cells, increasing their ability to secrete proinflammatory chemokines. *Journal of Medical Virology*, *67*, 528–533.

Caserta, M. ., Hall, C. ., Schnabel, K., Lofthus, G., Marino, A., Shelley, L., … Wang, H. (2010). Diagnostic assays for active infection with human herpesvirus 6 (HHV-6). *Journal of Clinical Virology*, *48*(1), 55–57.

Chan, P. ., Ng, H. K., Hui, M., & Cheng, A. F. (2001). Prevalence and distribution of human herpesvirus 6 variants A and B in adult human brain. *Journal of Medical*

Virology, *64*, 42–46.

Chang, C. ., Yu, K. ., Mbulaiteye, S. ., Hildesheim, A., & Bhatia, K. (2009). The extent of genetic diversity of Epstein-Barr virus and its geographic and disease patterns: a need for reappraisal. *Virus Research*, *143*, 209–221.

Chang, R. ., Fillingame, R. ., Paglieroni, T., & Glassy, F. . (1976). A procedure for quantifying the susceptibility of human lymphocytes to transformation by Epstein-Barr viruses. *Proceedings of the Society for Experimental Biology and Medicine*, *153*, 193–196.

Chen, H. ., Lung, M. ., Chan, K. ., Griffin, B. E., & Ng, M. . (1996). Tissue distribution of Epstein-Barr virus genotypes. *Journal of Virology*, *70*(10), 7301–7305.

Chen, J., Rowe, C. ., Jardetzky, T. ., & Longnecker, R. (2012). The KGD motif of Epstein-Barr virus gH/gL is bifunctional, orchestrating infection of B cells and epithelial cells. *mBio*, *3*(1), e00290-11.

Cicin-sain, L., Brien, J. D., Uhrlaub, J. L., Drabig, A., & Marandu, T. F. (2012). Cytomegalovirus Infection Impairs Immune Responses and Accentuates T-cell Pool Changes Observed in Mice with Aging, *8*(8). http://doi.org/10.1371/journal.ppat.1002849

Clark, D. a, Nacheva, E. P., Leong, H. N., Brazma, D., Li, Y. T., Tsao, E. H. F., … Griffiths, P. D. (2006). Transmission of integrated human herpesvirus 6 through stem cell transplantation: implications for laboratory diagnosis. *The Journal of Infectious Diseases*, *193*(7), 912–6. http://doi.org/10.1086/500838

Collot, S., Petit, B., Bordessoule, D., Alain, S., Touati, M., Denis, F., & Ranger-Rogez, S. (2002). Real-time PCR for quantification of human herpesvirus 6 DNA from lymph nodes and saliva. *Journal of Clinical Microbiology*, *40*, 2445–2451.

de Campos-Lima, P. O., Gavioli, R., Zhang, Q., Wallace, L. E., Dolcetti, R., Rowe, M., … Masucci, M. G. (1993). HLA-Al 1 Epitope Loss Isolates of Epstein-Barr Virus from a Highly A11+ Population. *Science*, *260*(April), 98–101.

Deback, C., Géli, J., Aït-Arkoub, Z., Angleraud, F., Gautheret-Dejean, A., Agut, H., & Boutolleau, D. (2009). Use of the

Roche LightCycler 480 system in a routine laboratory setting for molecular diagnosis of opportunistic viral infections: Evaluation on whole blood specimens and proficiency panels. *Journal of Virological Methods*, *159*(2), 291–294.

DeLorenze, G. ., Munger, K. ., Lennette, E. ., Orentreich, N., Vogelman, J. ., & Ascherio, A. (2006). Epstein-Barr virus and multiple sclerosis: evidence of association from a prospective study with long-term follow-up. *Archives of Neurology*, *63*, 839–844.

Demogines, A., Abraham, J., Choe, H., Farzan, M., & Sawyer, S. L. (2013). Dual Host-Virus Arms Races Shape an Essential Housekeeping Protein, *11*(5). http://doi.org/10.1371/journal.pbio.1001571

Depledge, D. P., Kundu, S., Jensen, N. ., Gray, E. ., Jones, M., Steinberg, S., … Breuer, J. (2014). Deep sequencing of viral genomes provides insight into the evolution and pathogenesis of varicella zoster virus and its vaccine in humans. *Molecular Biology and Evolution*, *31*(2), 397–409.

Depledge, D. P., Palser, A. L., Watson, S. J., Lai, I. Y.-C., Gray, E. R., Grant, P., … Breuer, J. (2011). Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. *PLoS ONE*, *6*(11), e27805. http://doi.org/10.1371/journal.pone.0027805

Di Luca, D., Mirandola, P., Ravaioli, T., Dolcetti, R., Frigatti, A., Bovenzi, P., … Cassal, E. (1995). Human herpesviruses 6 and 7 in salivary glands and shedding in saliva of healthy and human immunodeficiency virus positive individuals. *Journal of Medical Virology*, *45*, 462–468.

Dominguez, G., Dambaugh, T. R., Stamey, F. R., Dewhurst, S., Inoue, N., & Pellett, P. E. (1999). Human herpesvirus 6B genome sequence: coding content and comparison with human herpesvirus 6A. *Journal of Virology*, *73*(10), 8040–52. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=112820&tool=pmcentrez&rendertype=abstract

Donaldson, C. D., Clark, D. A., Kidd, I. M., Breuer, J., & Depledge, D. D. (2013). Genome sequence of human herpesvirus 7 strain UCL-1. *Genome Announcements*, *1*,

e00830-13.
Donati, D., Akhyani, N., Fodgell-Hann, A., Cermelli, C., Cassiani-Ingoni, R., Vortmeyer, A., … Jacobson, S. (2003). Detection of human herpesvirus-6 in mesial temporal lobe epilepsy surgical brain resections. *Neurology*, *61*, 1405–1411.
Dyson, P. ., & Farrell, P. . (1985). Chromatin structure of Epstein- Barr virus. *Journal of General Virology*, *66*, 1931–1940.
E.V, K., Senkevich, T. ., & Dolja, V. . (2006). The ancient Virus World and evolution of cells. *Biology Direct*, *1*(29).
Ebert, K., Depledge, D. P., Breuer, J., Harman, L., & Elliott, G. (2013). Mode of virus rescue determines the acquisition of VHS mutations in VP22-negative herpes simplex virus 1. *Journal of Virology*, *87*, 10389–10393.
Edwards, R. ., Seillier-Moiseiwitsch, F., & Raab-traub, N. (1999). Signature amino acid changes in latent membrane protein 1 distinguish Epstein-Barr virus strains. *Virology*, *261*(1), 79–95.
Ejrnaes, M., Filippi, C. ., Martinic, M. ., Ling, E. ., Togher, L. ., Crotty, S., & von Herrath, M. . (2006). Resolution of a chronic viral infection after interleukin-10 receptor blockade. *Journal of Experimental Medicine*, *203*(11), 24612472.
Enard, D., Cai, L., Gwennap, C., & Petrov, D. A. (2016). Viruses are a dominant driver of protein adaptation in mammals. *eLIFE*, *5*, e12469. http://doi.org/10.7554/eLife.12469
Faulkner, G. ., Krajewski, A. ., & Crawford, D. . (2000). The ins and outs of EBV infection. *Trends in Microbiology*, *8*, 185–189.
Flamand, L., Gosselin, J., D'addario, M., Hiscott, J., Ablashi, D. V., Gallo, R. ., & Menezes, J. (1991). Human herpesvirus 6 induces interleukin 1 beta and tumor necrosis factor alpha, but not interleukin 6, in peripheral blood mononuclear cell cultures. *Journal of Virology*, *65*, 5105–5110.
Flamand, L., Gosselin, J., Stefanescu, I., Ablashi, D. ., & Menezes, J. (1995). Immunosuppressive effect of human herpesvirus 6 on T-cell functions: suppression of interleukin 2 synthesis and cell proliferation. *Blood*, *85*,

1263–1271.
Flavell, K. ., & Murray, P. G. (2000). Hodgkin's disease and the Epstein-Barr virus. *Molecular Pathology*, *53*, 262–269.
Fox, J. D., Briggs, M., Ward, P. A., & Tedder, R. S. (1990). Human herpesvirus 6 in salivary glands. *Lancet (London, England)*, *336*, 590–593.
Fumagalli, M., Cagliani, R., Riva, S., Pozzoli, U., Biasin, M., Piacentini, L., … Sironi, M. (2010). Population Genetics of IFIH1 : Ancient Population Structure , Local Selection , and Implications for Susceptibility to Type 1 Diabetes Research article. *Molecular Biology and Evolution*, *27*(11), 2555–2566. http://doi.org/10.1093/molbev/msq141
Furlini, G., Vignoli, M., Ramazzotti, E., Re, M. C., Visani, G., & La, P. (996). A concurrent human herpesvirus-6 infection renders two human hematopoietic progenitor (TF-1 and KG-1) cell lines susceptible to human immunodeficiency virus type-1. *Blood*, *87*, 4737–4745.
Gao, Z., Krithivas, A., Finan, J. ., Semmes, O. ., Zhou, S., Wang, Y., & Hayward, S. . (1998). The Epstein-Barr virus lytic transactivator Zta interacts with the heli- case-primase replication proteins. *Journal of Virology*, *72*(11), 8559–8567.
Geoghegan, J. L. (2017). Comparative analysis estimates the relative frequencies of co-divergence and cross- species transmission within viral families. *PLoS Pathogens*, *13*(2), e1006215. http://doi.org/10.1371/journal.ppat.1006215
Gompels, U. A., Nicholas, J., Lawrence, G., Jones, M., Thomson, B. J., Martin, M. E., … Macaulay, H. A. (1995). The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. *Virology*, *209*(1), 29–51. http://doi.org/10.1006/viro.1995.1228
Gravel, A., Dubuc, I., Morissette, G., Sedlack, R. ., Jerome, K. ., & Flamand, L. (2015). Inherited chromosomally integrated human herpesvirus 6 as a predisposing risk factor for the development of angina pectoris. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(26), 8058–8063.
Greninger, A. ., Knudsen, G. ., Roychoudhury, P., Hanson, D.

., Sedlak, R. ., Xie, H., … Jerome, K. R. (n.d.). Genomic and proteomic analysis of Human herpesvirus 6 reveals distinct clustering of acute versus inherited forms and reannotation of reference strain. *https://www.biorxiv.org/content/early/2017/08/27/181248*.

Griffin, B. ., Gram, A. ., Mulder, A., Van Leeuwen, D., Claas, F. H. ., Wang, F., … Wiertz, E. (2013). Epstein-Barr virus BILF1 evolved to downregulate cell surface display of a wide range of HLA class I molecules through their cytoplasmic tail. *Journal of Immunology*, *190*(4), 1672–1684.

Griffiths, P. D., Ait-Khaled, M., Bearcroft, C. P., Clark, D. A., Quaglia, A., Davies, S. E., … Emery, V. C. (1999). Human herpesviruses 6 and 7 as potential pathogens after liver transplant: Prospective comparison with the effect of cytomegalovirus. *Journal of Medical Virology*, *59*(4), 496–501. http://doi.org/10.1002/(SICI)1096-9071(199912)59:4<496::AID-JMV12>3.0.CO;2-U

Grivel, J.-C., Santoro, F., Chen, S., Fagá, G., Malnati, M. S., Ito, Y., … Lusso, P. (2003). Pathogenic effects of human herpesvirus 6 in human lymphoid tissue ex vivo. *Journal of Virology*, *77*(15), 8280–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165251&tool=pmcentrez&rendertype=abstract

Grose, C. (2012). Pangaea and the Out-of-Africa Model of Varicella-Zoster Virus Evolution and Phylogeography. *Journal of Virology*, *86*(18), 9558–9565. http://doi.org/10.1128/JVI.00357-12

Gutierrez, M. ., Ibrahim, M. ., Dale, J. ., Greiner, T. ., Straus, S. ., & Bhatia, K. (2002). Discrete alterations in the BZLF1 promoter in tumor and non-tumor-associated Epstein-Barr virus. *Journal of National Cancer Institute*, *94*(23), 1757–1763.

Gutierrez, M. ., Kingma, D. ., Sorbara, L., Tran, M., Raffeld, M., Jaffe, E. ., … Bhatia, K. (2000). Association of EBV strains, defined by multiple loci analyses, in non-Hodgkin lymphomas and reactive tissues from HIV positive and HIV negative patients. *Leukemia & Lymphoma*, *37*(3–4), 425–429.

Gutierrez, M. ., Raj, A., Spangler, G., Sharma, A., Hussain, A., Judde, J. ., … Bhatia, K. (1997). Sequence variations

in EBNA-1 may dictate restriction of tissue distribution of Epstein- Barr virus in normal and tumour cells. *Journal of Genetic Virology*, *78*, 1663–1670.

Hadinoto, V., Shapiro, M., Sun, C. ., & Thorley-Lawson, D. . (2009). The Dynamics of EBV Shedding Implicate a Central Role for Epithelial Cells in Amplifying Viral Output. *PLoS Pathogens*, *5*(7), e1000496.

Hahn, G., Jores, R., & Mocarski, E. (1998). Cytomegalovirus remains latent in a common precursor of dendritic and myeloid cells. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(7), 3937–3942.

Halenius, A., Hauka, S., Dölken, L., Stindt, J., Reinhard, H., Wiek, C., … Hengel, H. (2011). Human Cytomegalovirus Disrupts the Major Histocompatibility Complex Class I Peptide-Loading Complex and Inhibits Tapasin Gene Transcription. *Journal of Virology*, *85*(7), 3473–3485.

Hall, C. B., Caserta, M. T., Schnabel, K., Shelley, L. M., Marino, A. S., Carnahan, J. A., … McDermott, M. P. (2008). Chromosomal Integration of Human Herpesvirus 6 Is the Major Mode of Congenital Human Herpesvirus 6 Infection. *Pediatrics*, *122*(3), 513–520. http://doi.org/10.1542/peds.2007-2838

Hall, C., Caserta, M. ., Schnabel, K. ., Boettrich, C., McDermott, M. ., Lofthus, G. ., … Dewhurst, S. (2004). Congenital infections with human herpesvirus 6 (hhv6) and human herpesvirus 7 (hhv7). *Journal of Pediatrics*, *145*(4), 472–477.

Hammerschmidt, W., & Sugden, B. (1988). Identification and characterization of oriLyt, a lytic origin of DNA replication of Epstein-Barr virus. *Cell*, *55*, 427–433.

Harma, M., Hockerstedt, K., & Launtenschlager, I. (2003). Human herpesvirus-6 and acute liver failure. *Transplantation*, *76*, 536–539.

Hasegawa, A., Yasukawa, M., Sakai, I., & Fujita, S. (2001). Transcriptional downregulation of CXC chemokine receptor 4 induced by impaired association of transcription regulator YY1 with c-myc in human herpesvirus 6-infected cells. *Journal of Immunology*, *166*, 1125–1131.

Henle, W., Diehl, V., Kohn, G., Zur Hausen, H., & Henle, G.

(1967). Herpes-type virus and chromosome marker in normal leukocytes after growth with irradiated Burkitt cells. *Science*, *157*, 1064–1065.

Hill, J. A., Hallsedlak, R., Magaret, A., Huang, M., Zerr, D. M., Jerome, K. R., … States, U. (2016). Efficient identification of inherited chromosomally integrated human herpesvirus 6 using specimen pooling, *77*, 71–76. http://doi.org/10.1016/j.jcv.2016.02.016.Efficient

Hirata, Y., Kondo, K., & Yamanishi, K. (2001). Human herpesvirus 6 downregulates major histocompatibility complex class I in dendritic cells. *Journal of Medical Virology*, *65*, 576–583.

Hjalgrim, H., Askling, J., Sorensen, P., Madsen, M., Rosdahl, N., Storm, H. ., … Melbye, M. (2000). Risk of Hodgkin's disease and other cancers after infectious mononucleosis. *Journal of Natural Cancer Institute*, *92*, 1522–1528.

Hong, E. ., & Park, J. . (2012). Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & Informatics*, *10*(2), 117–122.

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., … Tomonaga, K. (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature*, *463*, 84–87.

Hu, X., Margolis, H. ., Purcell, R. ., Ebert, J., & Robertson, B. . (2000). Identification of hepatitis b virus indigenous to chimpanzees. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(4), 1661–1664.

Huang, Y., Hidalgo-Bravo, A., Zhang, E., Cotton, V. E., Mendez-Bermudez, A., Wig, G., … Royle, N. J. (2014). Human telomeres that carry an integrated copy of human herpesvirus 6 are often short and unstable, facilitating release of the viral genome from the chromosome. *Nucleic Acids Research*, *42*(1), 315–27. http://doi.org/10.1093/nar/gkt840

Huh, J. (2012). Epidemiologic overview of malignant lymphoma. *The Korean Journal of Hematology*, *47*(2), 92–104. http://doi.org/10.5045/kjh.2012.47.2.92

Ihira, M., Yoshikawa, T., Ishii, J., Nomura, M., Hishida, H., Ohashi, M., … Asano, Y. (2002). Serological examination

of human herpesvirus 6 and 7 in patients with coronary artery disease. *Journal of Medical Virology*, *67*(4), 534–537. http://doi.org/10.1002/jmv.10134

Iizasa, H., Nanbo, A., Nishikawa, J., Jinushi, M., & Yoshiyama, H. (2012). Epstein-Barr virus (EBV)-associated gastric carcinoma. *Viruses*, *4*, 3420–3439.

Inagi, R., Guntapong, R., Nakao, M., Ishino, Y., Kawanishi, K., Isegawa, Y., & Yamanishi, K. (1996). Human herpesvirus 6 induces IL-8 gene expression in human hepatoma cell line, Hep G2. *Journal of Medical Virology*, *49*(1), 34–40.

Jarrett, R. F., Clark, D. A., Josephs, S. F., & Onions, D. E. (1990). Detection of human herpesvirus-6 DNA in peripheral blood and saliva. *Journal of Medical Virology*, *32*, 73–76.

Kakadia, M. ., Rybka, W. ., Stweart, J. ., Patton, J. ., Stamey, F. ., Elsawy, M., … Armstrong, J. . (1996). Human herpesvirus 6: infection and disease following autologous and allogeneic bone marrow transplantation. *Blood*, *87*(12), 5341–5354.

Kakimoto, M., Hasegawa, A., Fujita, S., & Yasukawa, M. (2002). Phenotypic and functional alterations of dendritic cells induced by human herpesvirus 6 infection. *Journal of Virology*, *76*(20), 10338–10345.

Kang, M., & Kieff, E. (2015). Epstein – Barr virus latent genes, *47*(1), e131-16. http://doi.org/10.1038/emm.2014.84

Kerns, J. A., Emerman, M., & Malik, H. S. (2008). Positive Selection and Increased Antiviral Activity Associated with the PARP-Containing Isoform of Human Zinc-Finger Antiviral Protein. *PLoS Genetics*, *4*(1), e21. http://doi.org/10.1371/journal.pgen.0040021

Kimura, M. (1977). preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, *267*, 275–276.

Kirchmaier, A. ., & Sugden, B. (1995). Plasmid maintenance of derivatives of oriP of Epstein-Barr virus. *Journal of Virology*, *69*, 1280–1283.

Kwei, K., Tang, X., Lok, A. ., Sureau, C., Garcia, T., Li, J., … Tong, S. (2013). Impaired Virion Secretion by Hepatitis B Virus Immune Escape Mutants and Its Rescue by Wild-

Type Envelope Proteins or a Second-Site Mutation. *Journal of Virology*, *87*(4), 2352–2357.

Kwok, H., Wu, C. ., Palser, A. L., Kellam, P., Sham, P. ., Kwong, D. ., & Chiang, A. . (2014). Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsy samples. *PLoS ONE*, *88*, 10662–10672.

Lavergne, a., Donato, D., Gessain, a., Niphuis, H., Nerrienet, E., Verschoor, E. J., & Lacoste, V. (2014). African Great Apes Are Naturally Infected with Roseoloviruses Closely Related to Human Herpesvirus 7. *Journal of Virology*, *88*(September), 13212–13220. http://doi.org/10.1128/JVI.01490-14

Leach, C. T., Newton, E. R., McParlin, S., & Jenson, H. B. (1994). Human herpesvirus 6 infection of the female genital tract. *Journal of Infectious Diseases*, *169*(6), 1281–1283.

Leong, H. M., Tuke, P. W., Tedder, R. S., Khanom, A. B., Eglin, R. P., Atkinson, C. E., … Clark, D. A. (2007). The prevalence of chromosomally integrated human herpesvirus 6 genomes in the blood of UK blood donors. *Journal of Medical Virology*, *79*, 45–51.

Levin, L. ., Munger, K. ., O'Reilly, E. ., Falk, K. ., & Ascherio, A. (2010). Primary infection with the Epstein-Barr virus and risk of multiple sclerosis. *Annals of Neurology*, *67*, 824–830.

Ling, P. D., Lednicky, J. A., Keitel, W. A., Poston, D. G., White, Z. S., Peng, R., … Butel, J. S. (2003). The Dynamics of Herpesvirus and Polyomavirus Reactivation and Shedding in Healthy Adults : A 14-Month Longitudinal Study, *187*, 1571–1580.

Linhares, M. I., Eizuru, Y., Tateno, S., & Minamishima, Y. (1991). Seroprevalence of human herpesvirus 6 infection in Brazilian and Japanese populations in the north-east of Brazil. *Microbiology and Immunology*, *35*(11), 1023–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1663574

Liu, H., Wang, Y., Liao, C., & Kuang, Y. (2005). Adaptive evolution of primate TRIM5 α , a gene restricting HIV-1 infection ☆. *Gene*, *362*, 109–116.

http://doi.org/10.1016/j.gene.2005.06.045
Ljungman, P. (2002). Beta-herpesvirus challenges in the transplant recipient. *Journal of Infectious Diseases*, *186*(Suppl 1), S99–S109.
Longnecker, R., Kieff, E., & Cohen, J. (2013). *Epstein-Barr virus* (6th ed.). Philadelphia, PA: Lippincott/The Williams and Wilkins Co.
Lucchesi, W., Brandy, G., Dittrich-Breiholz, O., Kracht, M., Russ, R., & Farrell, P. . (2008). Differential gene regulation by Epstein-Barr virus type 1 and type 2 EBNA2. *Journal of Virology*, *82*(15), 7456–7466.
Luka, J., Okano, M., & Thiele, G. (1990). Isolation of human herpesvirus-6 from clinical specimens using human fibroblast cultures. *Journal of Clinical Laboratory Analysis*, *4*(6), 483–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2178187
Luppi, M., Barozzi, P., Morris, C., Garber, R. L., Bonacorsi, G., Donello, A., … Torelli, G. (1999). Human herpesvirus 6 latently infects early bone marrow progenitors in vivo. *Journal of Virology*, *73*(1), 754–759.
Luppi, M., Marasca, R., Barozzi, P., Ferrari, S., Ceccherini-Nelli, L., Batoni, G., … Torelli, G. (1993). Three cases of human herpesvirus-6 latent infection: integration of viral genome in peripheral blood mononuclear cell DNA. *Journal of Medical Virology*, *40*(1), 44–52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8099945
Lusso, P., Malnati, M., De Maria, A., Balotta, C., DeRocco, S. E., Markham, P. D., & Gallo, R. C. (1991). Productive infection of CD4+ and CD8+ mature human T cell populations and clones by human herpesvirus 6. *Journal of Immunology*, *147*(2), 685–691.
Lusso, P., Markham, P. D., Tschachler, E., di Marzo Veronese, F., Salahuddin, S. Z., Ablashi, D. V, … Gallo, R. C. (1988). In vitro cellular tropism of human B-lymphotropic virus (human herpesvirus-6). *The Journal of Experimental Medicine*, *167*(5), 1659–70. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3259254
Lythgoe, K. ., Gardner, A., Pybus, O. ., & Grove, J. (2017). Short-Sighted Virus Evolution and a Germline Hypothesis for Chronic Viral Infections. *Trends in Microbiology*, *25*(5), 336–348.

Maeda, A., Sata, T., Enzan, H., Tanaka, K., Wakiguchi, H., Kurashige, T., … Kurata, T. (1993). The evidence of human herpesvirus 6 infection in the lymphnodes of Hodgkin's disease. *Virchows Archiv A. Pathol. Anat.*, *423*(1), 71–75. http://doi.org/10.1007/BF01606435

Mahmoud, N. F., Kawabata, A., Tang, H., Wakata, A., Wang, B., Serada, S., … Mori, Y. (2016). Human herpesvirus 6 U11 protein is critical for virus infection. *Virology*, *489*, 151–157. http://doi.org/10.1016/j.virol.2015.12.011

Mandage, R., Telford, M., Rodrı, J. A., Farre, X., Laayouni, H., Marigorta, U. M., … Santpere, G. (2017). Genetic factors affecting EBV copy number in lymphoblastoid cell lines derived from the 1000 Genome Project samples. *PloS One*, *12*(6), e0179446.

Matteo, C., Manzari, C., Lionetti, C., Mechelli, R., Anastasiadou, E., Buscarinu, M. C., … Horner, D. S. (2016). Geographic Population Structure in Epstein-Barr Virus Revealed by Comparative Genomics, *8*(2015), 3284–3291. http://doi.org/10.1093/gbe/evw226

Mcgeoch, D. J., & Gatherer, D. (2005). Integrating Reptilian Herpesviruses into the Family Herpesviridae. *Journal of Virology*, *79*(2), 725–731. http://doi.org/10.1128/JVI.79.2.725

Mcgeoch, D. J., Rixon, F. J., & Davison, A. J. (2006). Topics in herpesvirus genomics and evolution. *Virus Research*, *117*, 90–104. http://doi.org/10.1016/j.virusres.2006.01.002

Melbye, M., Ebbesen, P., Levine, P. ., & Bennike, T. (1984). Early primary infection and high Epstein- Barr virus antibody titers in Greenland Eskimos at high risk for nasopharyngeal carcinoma. *International Journal of Cancer*, *34*, 619–623.

Meyer, M., & Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. http://doi.org/10.1101/pdb.prot5448

Meyne, J., Ratliff, R. L., & Moyzis, R. K. (1989). Conservation of the human telomere sequence (TTAGGG) among vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*,

86(September), 7049–7053.
Miller, G., Lisco, H., Kohn, H. ., & Stitt, D. (1971). Establishment of cell lines from normal adult human blood leukocytes by exposure to Epstein Barr virus and neutralization by human sera with Epstein-Barr virus antibody. *Proceedings of the Society for Experimental Biology and Medicine*, *137*, 1459–1465.
Miller, N., & Hutt-fletcher, L. M. (1992). Epstein-Barr virus enters B cells and epithelial cells by different routes. *Journal of Virology*, *66*, 3409–3414.
Milne, R. ., Mattick, C., Nicholson, L., Devaraj, P., Alcami, A., & Gompels, U. A. (2000). ANTES binding and downregulation by a novel human herpesvirus 6 beta chemokine receptor. *Journal of Immunology*, *164*, 2396–2404.
Mori, Y., Yang, X., Akkapaiboom, P., Okuno, T., & Yamanishi, K. (2003). Human herpesvirus 6 variant A glycoprotein H-glycoprotein L-glycoprotein Q complex associates with human CD46. *Journal of Virology*, *77*, 4992–4999.
Morissette, G., & Flamand, L. (2010a). Herpesviruses and chromosomal integration. *Journal of Virology*, *84*(23), 12100–12109. http://doi.org/10.1128/JVI.01169-10
Morissette, G., & Flamand, L. (2010b). Herpesviruses and chromosomal integration. *Journal of Virology*, *84*(23), 12100–9. http://doi.org/10.1128/JVI.01169-10
Münz, C. (2015). *Epstein-Barr virus volume 1. One herpes virus: many diseases*. Zurich: Springer.
Nacheva, E. P., Ward, K. N., Brazma, D., Virgili, A., Howard, J., Leong, H. N., & Clark, D. A. (2008). Human Herpesvirus 6 Integrates Within Telomeric Regions as Evidenced by Five Different Chromosomal Sites. *Journal of Medical Virology*, *80*, 1952–1958.
Nemerow, G. ., & Cooper, N. . (1984). Early events in the infection of human B lymphocytes by Epstein-Barr virus: the internalization process. *Virology*, *132*, 186–198.
Nielsen, L., & Vestergaard, B. F. (1996). Competitive ELISA for detection of HHV-6 antibody: seroprevalence in a danish population. *Journal of Virological Methods*, *56*(2), 221–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8882652
Nielsen, R., & Yang, Z. (2003). Estimating the Distribution of

Selection Coefficients from Phylogenetic Data with Applications to Mitochondrial and Viral DNA, *20*, 1231–1239. http://doi.org/10.1093/molbev/msg147

Okuno, T., Takahashi, K., Balachandra, K., Shiraki, K., Yamanishi, K., Takahashi, M., & Baba, K. (1989). Seroepidemiology of human herpesvirus 6 infection in normal children and adults. *Journal of Clinical Microbiology*, *27*(4), 651–3. http://doi.org/10.1016/0888-0786(96)87294-5

Olsson, J., WIkby, A., Johansson, B., Lofgren, S., Nilsson, B. ., & Ferguson, F. . (2000). Age-related change in peripheral blood T-lymphocyte subpopulations and cytomegalovirus infection in the very old: the Swedish longitudinal OCTO immune study. *Mechanisms of Ageing and Development*, *121*(1–3), 187–201.

Packham, G., Economou, A., Rooney, C. ., Rowe, D. T., & Farrell, P. . (1990). Struc- ture and function of the Epstein-Barr virus BZLF1 protein. *Journal of Virology*, *64*(5), 2110–2116.

Pakpoor, J., Disanto, G., Gerber, J. ., Dobson, R., Meier, U. ., Giovannoni, G., & Ramagopalan, S. . (2013). The risk of developing multiple sclerosis in individuals seronegative for Epstein-Barr virus: a meta-analysis. *Multiple Sclerosis*, *19*, 162–166.

Palser, A. L., Grayson, N. E., White, R. E., Corton, C., Correia, S., Ba abdullah, M. M., … Kellam, P. (2015). Genome Diversity of Epstein-Barr Virus from Multiple Tumor Types and Normal Infection. *Journal of Virology*, *89*(10), 5222–5237. http://doi.org/10.1128/JVI.03614-14

Parker, B. ., Bankier, A., Satchwell, S., Barrell, B., & Farrell, P. . (1990). Sequence and transcription of Raji Epstein-Barr virus DNA spanning the B95–8 deletion region. *Virology*, *179*(1), 339–346.

Pass, R. (2001). Cytomegalovirus. In L. W. and W. Co. (Ed.), *Field virology 4th Ed.* (4th ed., pp. 2675–2705). Philadelphia, PA: Lippincott/The Williams and Wilkins Co.

Patel, M. ., Emerman, M., & Malik, H. . (2011). Paleovirology — ghosts and gifts of viruses past. *Current Opinions in Virology*, *1*(4), 304–309.

Pathmanathan, R., Prasad, U., Sadler, R., Flynn, K., & Raab-Traub, N. (1995). Clonal proliferations of cells infected

with Epstein-Barr virus in preinvasive lesions related to nasopharyngeal carci- noma. *New England Journal of Medicine*, *333*, 693–698.

Pawaelec, G., Akbar, A., Caruso, C., Effros, R., Grubeck-Loebenstein, B., & Wikby, A. (2004). Is immunosenescence infectious? *Trends of Immonology*, *25*(8), 406–410.

Pfeffer, S., Zavolan, M., Grässer, F. ., Chien, M., Russo, J. ., Ju, J., … Tuschl, T. (2004). Identification of virus-encoded microRNAs. *Science*, *304*, 734–736.

Pfeiffer, B., Berneman, Z. W. I. N., Neipel, F., Chang, C. K., Tirwatnapong, S., & Chandran, B. (1993). Identification and Mapping of the Gene Encoding the Glycoprotein Complex gp82-gplO5 of Human Herpesvirus 6 and Mapping of the Neutralizing Epitope Recognized by Monoclonal Antibodies, *67*(8), 4611–4620.

Pfeiffer, B., & Thomson, B. (1995). Identification and Characterization of a cDNA Derived from Multiple Splicing That Encodes Envelope Glycoprotein gp105 of Human Herpesvirus 6, *69*(6), 3490–3500.

Pimenoff, V. N., Oliveira, C. M. De, & Bravo, I. G. (2016). Transmission between Archaic and Modern Human Ancestors during the Evolution of the Oncogenic Human Papillomavirus. *Molecular Biology and Evolution*, *34*(1), 4–19. http://doi.org/10.1093/molbev/msw214

Piriou, E., Asito, A. ., Sumba, P. ., Fiore, N., Middeldorp, J. ., Moorman, A. ., … Rochford, R. (2012). Early age at time of primary Epstein-Barr virus infection results in poorly controlled viral infection in infants from Western Kenya: clues to the etiology of endemic Burkitt lymphoma. *Journal of Infectious Diseases*, *205*, 906–913.

Pohl, D., Krone, B., Rostasy, K., Kahler, E., Brunner, E., Lehnert, M., … Henefeld, F. (2006). High seroprevalence of Epstein-Barr virus in children with multiple sclerosis. *Neurology*, *67*, 2063–2065.

Politou, M., Koutras, D., Kaparos, G., Valsami, S., Pittaras, T., Logothetis, E., … Kouskouni, E. (2014). Seroprevalence of HHV-6 and HHV-8 among blood donors in Greece. *Virology Journal*, *11*, 153. http://doi.org/10.1186/1743-422X-11-153

Raab-Traub, N. (2002). Epstein-Barr virus in the

pathogenesis of NPC. *Seminars in Cancer Biology*, *12*, 431–441.

Raab-Traub, N., & Flynn, K. (1986). The structure of the termini of the Epstein-Barr virus as a marker of clonal cellular proliferation. *Cell*, *47*, 883–889.

Rasaiyaah, J., Tan, C. P., Fletcher, A. J., Price, A. J., Blondeau, C., Hilditch, L., … Towers, G. J. (2013). HIV-1 evades innate immune recognition through specific cofactor recruitment. *Nature*, *503*(7476), 402–405. http://doi.org/10.1038/nature12769

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer, *29*(1), 24–26. http://doi.org/10.1038/nbt0111-24

Rohland, N., & Reich, D. (2012). Cost-effective , high-throughput DNA sequencing libraries for multiplexed target capture, 939–946. http://doi.org/10.1101/gr.128124.111.22

Roush, K. S., Domiati-Saad, R. K., Margraf, L. R., Krisher, K., Scheuermann, R. H., Rogers, B. B., & Dawson, D. B. (2001). Prevalence and cellular reservoir of latent human herpesvirus 6 in tonsillar lymphoid tissue. *American Journal of Clinical Pathology*, *166*, 648–654.

Rowe, M., Young, L. ., Cadwallader, K., Petti, L., Kieff, E., & Rickinson, A. . (1989). Distinction between Epstein-Barr virus type A (EBNA 2A) and type B (EBNA 2B) isolates extends to the EBNA 3 family of nuclear proteins. *Journal of Virology*, *63*(3), 1031–1039.

Sacaze, C., Henry, S., Icart, J., & Mariame, B. (2001). Tissue specific distribution of Epstein-Barr virus (EBV) BZLF1 gene variants in nasopharyngeal carcinoma (NPC) bearing patients. *Virus Research*, *81*(1–2), 133–142.

Salahuddin, S. Z., Ablashi, D. V, Markham, P. D., Josephs, S. F., Sturzenegger, S., Kaplan, M., … Gallo, R. C. (1986). Isolation of a New Virus , HBLV , in Patients with Lymphoproliferative Disorders, *234*, 596–601.

Sample, J., Young, L., Martin, B., Chatman, T., Kieff, E., Rickinsons, A., & Kieff, E. (1990). Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA-3B, and EBNA-3C genes. *Journal of Virology*, *64*(9), 4084–4092.

Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., &

Belshaw, R. (2010). Viral Mutation Rates. *Journal of Virology*, *84*(19), 9733–9748. http://doi.org/10.1128/JVI.00694-10

Santoro, F., Kennedy, P. E., Locatelli, G., Mainati, M. S., Berger, E. A., & Lusso, P. (1999). CD46 is a cellular receptor for human herpesvirus 6. *Cell*, *99*(7), 817–827.

Santpere, G., Darre, F., Blanco, S., Alcami, A., Villoslada, P., Mar Alba, M., & Navarro, A. (2014). Genome-Wide Analysis of Wild-Type Epstein-Barr Virus Genomes Derived from Healthy Individuals of the 1000 Genomes Project. *Genome Biology and Evolution*, *6*(4), 846–860. http://doi.org/10.1093/gbe/evu054

Sawyer, S. L., Emerman, M., & Malik, H. S. (2004). Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G. *PLoS Biology*, *2*(9), e275. http://doi.org/10.1371/journal.pbio.0020275

Sawyer, S. L., Emerman, M., & Malik, H. S. (2007). Discordant Evolution of the Adjacent Antiretroviral Genes TRIM22 and TRIM5 in Mammals. *PLoS Pathogens*, *3*(12), e197. http://doi.org/10.1371/journal.ppat.0030197

Sawyer, S. L., Wu, L. I., Emerman, M., & Malik, H. S. (2005). Positive selection of primate TRIM5 identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(8), 2832–2837.

Saxinger, C., Polesky, H., Eby, N., Grufferman, S., Murphy, R., Tegtmeir, G., … Hung, C. (1988). Antibody reactivity with HBLV (HHV-6) in U.S. populations. *Journal of Virological Methods*, *21*, 199–208.

Schirmer, E. C., Wyatt, L. S., Yamanishi, K., Rodriguez, W. J., & Frenkel, N. (1991). Differentiation between two distinct classes of viruses now classified as human herpesvirus 6. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(13), 5922–6. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51990&tool=pmcentrez&rendertype=abstract

Sculley, T. ., Apolloni, A., Stumm, R., Moss, D. ., Mueller-Lantczh, N., Misko, I. ., & Cooper, D. . (1989). Expression of Epstein-Barr virus nuclear antigens 3, 4, and 6 are altered in cell lines contain- ing B-type virus.

Virology, *171*, 401–408.
Sedlak, R. H., Cook, L., Huang, M., Magaret, A., Zerr, D. M., Boeckh, M., & Jerome, K. R. (2014). Identification of Chromosomally Integrated Human Herpesvirus 6 by Droplet Digital. *Clinical Chemistry*, *60*, 765–772. http://doi.org/10.1373/clinchem.2013.217240
Seki, S., & Matano, T. (2012). CTL Escape and Viral Fitness in HIV/SIV Infection. *Frontiers in Microbiology*, *2*.
Serafini, B., Severa, M., Columba-Cabezas, S., Rosicarelli, B., Veroni, C., Chiappetta, G., … Aloisi, F. (2010). Epstein-Barr Virus Latent Infection and BAFF Expression in B Cells in the Multiple Sclerosis Brain: Implications for Viral Persistence and Intrathecal B-Cell Activation. *Journal of Neuropathology & Experimental Neurology*, *69*(7), 677–693.
Shahani, L. (2014). HHV-6 encephalitis presenting as status epilepticus in an immunocompetent patient. *BMJ Case Reports*, *2014*. http://doi.org/10.1136/bcr-2014-205880
Shannon-Lowe, C. ., Nauhierl, B., Baldwin, G., Rickinson, A. ., & Delecluse, H. (2006). Resting B cells as a transfer vehicle for Epstein–Barr virus infection of epithelial cells. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(18), 7065–7070.
Siddon, A., Lozovatsky, L., Mohamed, A., & Hudnall, S. D. (2012). Human herpesvirus 6 positive Reed-Sternberg cells in nodular sclerosis Hodgkin lymphoma. *British Journal of Haematology*, *158*(5), 635–643. http://doi.org/10.1111/j.1365-2141.2012.09206.x
Sironi, M., Biasin, M., Cagliani, R., Luca, M. De, Saulle, I., Lo, S., … Clerici, M. (2012). A Common Polymorphism in TLR3 Confers Natural Resistance to HIV-1 Infection. *Journal of Immunology*, *188*, 818–823. http://doi.org/10.4049/jimmunol.1102179
Sitki-green, D., Covington, M., & Raab-Traub, N. (2003). Compartmentalization and transmission of multiple epstein- barr virus strains in asymptomatic carriers. *Journal of Virology*, *77*(3), 1840–1877.
Slyker, J. ., Casper, C., Tapia, K., Richardson, B., Bunts, L., Huang, M. ., … John-Stewart, G. (2013). Clinical and virologic manifestations of primary Epstein- Barr virus (EBV) infection in Kenyan infants born to HIV-infected

women. *Journal of Infectious Diseases*, *207*, 1798–1806.
Smith, A., Santoro, F., Di Iullo, G., Dagna, L., Verani, A., & Lusso, P. (2003). Selective suppression of IL-12 production by human herpesvirus 6. *Blood*, *102*(8), 2877–2884.
Sola, P., Merelli, E., Marasca, R., Poggi, M., Luppi, M., Montorsi, M., & Torelli, G. (1993). Human herpesvirus 6 and multiple sclerosis: survey of anti-HHV-6 antibodies by immunofluorescence analysis and of viral sequences by polymerase chain reaction. *Journal of Neurology, Neurosurgery, and Psychiatry*, *56*(8), 917–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8394408
Soldan, S. S., Berti, R., Salem, N., Secchiero, P., Flamand, L., Calabresi, P. A., … Jacobson, S. (1997). Association of human herpes virus 6 (HHV-6) with multiple sclerosis: increased IgM response to HHV-6 early antigen and detection of serum HHV-6 DNA. *Nature Medicine*, *3*(12), 1394–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9396611
Sprunt, T. ., & Evans, F. . (1920). Mononuclear leucocytosis in reaction to acute infections ("infectious mononucleosis"). *Johns Hopkins Hospital Bulletin*, *31*, 410–417.
Strenger, V., Aberle, S. W., Nacheva, E. P., & Urban, C. (2013). Chromosomal integration of the HHV-6 genome in a patient with nodular sclerosis Hodgkin lymphoma. *British Journal of Haematology*, *161*(4), 594–595. http://doi.org/10.1111/bjh.12257
Sullivan, B. M., & Coscoy, L. (2008). Downregulation of the T-Cell Receptor Complex and Impairment of T-Cell Activation by Human Herpesvirus 6 U24 Protein □, *82*(2), 602–608. http://doi.org/10.1128/JVI.01571-07
Sullivan, B. M., & Coscoy, L. (2010). The U24 Protein from Human Herpesvirus 6 and 7 Affects Endocytic Recycling □, *84*(3), 1265–1275. http://doi.org/10.1128/JVI.01775-09
Takahashi, K., Sonoda, S., Higashi, K., Kondo, T., Takahashi, H., Takahashi, M., & Yamanishi, K. (1989). Predominant CD4 T-lymphocyte tropism of human herpesvirus 6-related virus. *Journal of Virology*, *63*(7), 3161–3. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/2542623

Tanaka-Taya, K., Sashihara, J., Kurahashi, H., Amo, K., Miyagawa, H., Kondo, K., … Yamanishi, K. (2004). Human herpesvirus 6 (HHV-6) is transmitted from parent to child in an integrated form and characterization of cases with chromosomally integrated HHV-6 DNA. *Journal of Medical Virology*, *73*(3), 465–73. http://doi.org/10.1002/jmv.20113

Thomson, B. J., Dewhurst, S., & Gray, D. (1994). Structure and Heterogeneity of the a Sequences of Human Herpesvirus 6 Strain Variants U1102 and Z29 and Identification of Human Telomeric Repeat Sequences at the Genomic Termini, *68*(5), 3007–3014.

Tolfvenstam, T., Enbom, M., Ghebrekidan, H., Rudén, U., Linde, A., Grandien, M., & Wahren, B. (2000). Seroprevalence of viral childhood infections in Eritrea. *Journal of Clinical Virology : The Official Publication of the Pan American Society for Clinical Virology*, *16*(1), 49–54. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10680740

Triantos, D., Leao, J. ., Porter, S. ., Scully, C. ., & Teo, C. . (1998). Tissue distribution of Epstein-Barr virus genotypes in hosts coinfected by HIV. *AIDS*, *12*(16), 2141–2146.

Tsai, M.-H., Raykova, A., Klinke, O., Bernhardt, K., Gärtner, K., Leung, C. S., … Delecluse, H.-J. (2013). Spontaneous Lytic Replication and Epitheliotropism Define an Epstein-Barr Virus Strain Found in Carcinomas. *Cell Reports*, *5*(2), 458–470. http://doi.org/10.1016/j.celrep.2013.09.012

Tugizov, S., Herrera, R., Veluppillai, P., Greespan, J., Greenspan, D., & Palefsky, J. . (2007). Epstein-Barr virus (EBV)-infected monocytes facilitate dissemination of EBV within the oral mucosal epithelium. *Journal of Virology*, *81*(11), 5484–5496.

Tweedy, J., Spyrou, M. A., Hubacek, P., Kuhl, U., Lassner, D., & Gompels, U. A. (2015). Analyses of germline, chromosomally integrated human herpesvirus 6A and B genomes indicate emergent infection and new inflammatory mediators. *The Journal of General Virology*, *96*(Pt 2), 370–89. http://doi.org/10.1099/vir.0.068536-0

Tweedy, J., Spyrou, M. A., Pearson, M., Lassner, D., Kuhl, U., & Gompels, U. A. (2016). Complete genome sequence of germline chromosomally integrated human herpesvirus 6A and analyses integration sites define a new human endogenous virus with potential to reactivate as an emerging infection. *Viruses*, *8*(1). http://doi.org/10.3390/v8010019

Uebelhoer, L., Han, J., Callendret, B., Mateu, G., Shoukry, N. ., Hanson, H. ., … Grakoui, A. (2008). Stable Cytotoxic T Cell Escape Mutation in Hepatitis C Virus Is Linked to Maintenance of Viral Fitness. *PLoS Pathogens*, *4*(9), e1000143.

Uppal, T., Banarjee, S., Sun, Z., Verma, S. ., & Robertson, E. . (2014). KSHV LANA—The Master Regulator of KSHV Latency. *Viruses*, *6*(12), 4961–4998.

van de Berg, P. ., Griffiths, S. ., Yong, S. ., Macaulay, R., Bemelman, F. ., Jackson, S., … van Lier, R. . (2010). Cytomegalovirus infection reduces telomere length of the circulating T cell pool. *Journal of Immunology*, *184*(7), 3417–3423.

Vasseur, E., Patin, E., Laval, G., Pajon, S., Fornarino, S., Crouau-roy, B., & Quintana-murci, L. (2011). The selective footprints of viral pressures at the human RIG-I-like receptor family. *Human Molecular Genetics*, *20*(22), 4462–4474. http://doi.org/10.1093/hmg/ddr377

Virgin, H. ., Wherry, E. ., & Ahmed, R. (2009). Redefining chronic viral infection. *Cell*, *138*, 30–50.

Walling, D. ., Brown, A. ., Etienne, W., Keitel, W. ., & Ling, P. . (2003). Multiple Epstein-Barr virus infections in healthy individuals. *Journal of Virology*, *77*(11), 6546–6550.

Walling, D. ., & Raab-Traub, N. (1994). Epstein-Barr virus intrastrain recombination in oral hairy leukoplakia. *Journal of Virology*, *68*(12), 7909–7917.

Wang, C., Zhao, L., & Lu, S. (2015). Role of terra in telomere length. *International Journal of Biological Sciences*, *11*, 316–323.

Wang, G. ., Kao, W. ., Murakami, P., Xue, Q. ., Chiou, R. ., Detrick, B., … Fried, L. . (2010). Cytomegalovirus infection and the risk of mortality and frailty in older women: a prospective observational cohort study. *American Journal of Epidemiology*, *171*(10), 1144–1152.

Wang, X., Kenyon, W. ., Li, Q., Mullberg, J., & Hutt-fletcher, L. M. (1998). Epstein-Barr virus uses different complexes of glycoproteins gH and gL to infect B lymphocytes and epithelial cells. *Journal of Virology*, *72*, 5552–5558.

Ward, K. N., Leong, H. N., Nacheva, E. P., Howard, J., Atkinson, C. E., Davies, N. W. S., … Clark, D. A. (2006). Human herpesvirus 6 chromosomal integration in immunocompetent patients results in high levels of viral DNA in blood, sera, and hair follicles. *Journal of Clinical Microbiology*, *44*(4), 1571–4. http://doi.org/10.1128/JCM.44.4.1571-1574.2006

WIkby, A., Johansson, B., Olsson, J., Lofgren, S., Nilsson, B. ., & Ferguson, F. . (2002). Expansions of peripheral blood CD8 T-lymphocyte subpopulations and an association with cytomegalovirus seropositivity in the elderly: the Swedish NONA immune study. *Exploring Gerontology*, *37*(2–3), 445–453.

Wu, Z., Mu, G., & Wang, L. (1997). Seroprevalence of human herpesvirus-6 in healthy population in two provinces of north China. *Chinese Medical Sciences Journal = Chung-Kuo I Hsueh K'o Hsueh Tsa Chih*, *12*(2), 111–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11324495

Wyatt, L. S., Balachandran, N., & Frenkel, N. (1990). Variations in the Replication and Antigenic Properties of Human Herpesvirus 6 Strains. *The Journal of Infectious Diseases*, *162*, 852–857.

Yamanishi K., Okuno T., Shiraki K., Takahashi M., Kondo T., Asano Y., K. T. (1988). Identification Of Human Herpesvirus-6 As a Causal Agent for Exanthem Subitum, (May), 1065–1067.

Yang, Z., & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation, *15*(12), 496–503.

Yao, K., Honarmand, S., Espinosa, A., Akhyani, N., Glaser, C., & Jacobson, S. (2009). Detection of human herpesvirus-6 in cerebrospinal fluid of patients with encephalitis. *Annals of Neurology*, *65*(3), 257–267. http://doi.org/10.1002/ana.21611

Yao, Q. Y., & Rickinson, A. B. (1985). A re-examination of the Epstein-Barr virus carrier state in healthy seropositive indiviudals, *2*.

Yates, J. ., & Guan, N. (1991). Epstein-Barr virus-derived plas- mids replicate only once per cell cycle and are not amplified after entry into cells. *Journal of Virology*, *65*, 483–488.

Yates, J., Warren, N., Reisman, D., & Sugden, B. (1984). A cis-acting ele- ment from the Epstein-Barr viral genome that per- mits stable replication of recombinant plasmids in latently infected cells. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(12), 3806–3810.

Yates, J., Warren, N., & Sugden, B. (1985). Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells. *Nature*, *313*, 812–815.

Yeo, W. M., Isegawa, Y., Chow, V. T. K., & Irol, J. V. (2008). The U95 Protein of Human Herpesvirus 6B Interacts with Human GRIM-19 : Silencing of U95 Expression Reduces Viral Load and Abrogates Loss of Mitochondrial Membrane Potential , *82*(2), 1011–1020. http://doi.org/10.1128/JVI.01156-07

Yoshida, R., Imai, T., Hieshima, K., Kusuda, J., Baba, M., Kitaura, M., … Yoshie, O. (1997). Molecular cloning of a novel human CC chemokine EBI1-ligand chemokine that is a specific functional ligand for EBI1, CCR7. *Journal of Biological Chemistry*, *272*(21), 13803–13809.

Yoshikawa, T., Asano, Y., Akimoto, S., Ozaki, T., Iwasaki, T., Kurata, T., … Nishiyama, Y. (2002). Latent infection of human herpesvirus 6 in astrocytoma cell line and alteration of cytokine synthesis. *Journal of Medical Virology*, *66*(4), 497–505.

Yoshikawa, T., Asano, Y., Ihira, M., Suzuki, K., Ohashi, M., Suga, S., … Nishiyama, Y. (2002). Human herpesvirus 6 viremia in bone marrow transplant recipients: clinical features and risk factors. *Blood*, *185*, 847–853.

Yoshikawa, T., Ohashi, M., Miyake, F., Fujita, A., Usui, C., Sugata, K., … Asano, Y. (2009). Exanthem Subitum-Associated Encephalitis: Nationwide Survey in Japan. *Pediatric Neurology*, *41*(5), 353–358. http://doi.org/10.1016/j.pediatrneurol.2009.05.012

Young, L. ., & Rickinson, A. B. (2004). Epstein-Barr virus: 40 years on. *Nature Reviews Cancer*, *4*, 757–768.

Zaiac, A. ., Blattman, J. ., Murali-Krishna, K., Sourdive, D. .,

Suresh, M., Altman, J. ., & Ahmed, R. (1998). Viral Immune Evasion Due to Persistence of Activated T Cells Without Effector Function. *Jpurnal of Experimental Medicine*, *188*(12), 2205–2213.

Zhang, E., Bell, A. J., Wilkie, G. S., Suárez, N. ., Batini, C., Veal, C. ., … Royle, N. J. (2017). Inherited chromosomally integrated human herpesvirus 6 genomes are ancient, intact and potentially able to reactivate from telomeres. *Journal of Virology*, *44*(August). http://doi.org/10.1128/JVI.01137-17

Zhao, J., Fan, H., Mu, G., Shen, X., & Cheng, X. (1997). Detection of human herpesvirus 6 (HHV-6) DNA in salivary glands by the polymerase chain reaction. *Chinese Medical Sciences Journal = Chung-Kuo I Hsueh K'o Hsueh Tsa Chih*, *12*, 126–128.

Zhou, Y., Chang, C. K., Qian, G., Chandran, B., & Wood, C. (1994). trans-activation of the HIV promoter by a cDNA and its genomic clones of human herpesvirus-6. *Virology*, *199*(2), 311–322.

Zimber, U., Adldinger, H. ., Lenoir, G. ., Vuillaume, M., Knebel-Doeberitz, M. ., Laux, G., … Bornkamm, G. . (1986). Geographical prevalence of two types of Epstein–Barr virus. *Virology*, *154*, 56–66.