UNIVERSITY COLLEGE LONDON

DOCTORAL THESIS

Genomic studies on the impact of host/virus interaction in EBV infection using massively parallel high throughput sequencing

Author:
Fanny WEGNER

Supervisors:
Prof. Judith BREUER
Prof. Benjamin CHAIN

A thesis submitted for the degree of Doctor of Philosophy

in the

Division of Infection & Immunity

2017

Declaration of Authorship

I, Fanny WEGNER, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Date:

UNIVERSITY COLLEGE LONDON

Abstract

Division of Infection & Immunity

Doctor of Philosophy

Genomic studies on the impact of host/virus interaction in EBV infection using massively parallel high throughput sequencing

by Fanny WEGNER

Epstein-Barr virus is one of the most common viral infections in humans and, once acquired, persists within its host throughout their life. EBV therefore represents an extremely successful virus, having evolved complex strategies to evade the host's innate and adaptive immune response during both initial and persistent stages of infection. While infection is mostly harmless in the majority of cases, EBV has the ability to be oncogenic in some individuals, and is associated with a wide range of malignancies as well as non-cancerous diseases.

To generate new and useful insights into the evolution of EBV interactions with its host, a hybridization-based target enrichment methodology was optimised to enable whole genome sequencing of EBV directly from clinical samples. This allowed the generation of whole genome sequences of EBV directly from blood for the first time.

This methodology was subsequently applied to a number of distinct EBV sample collections and the resulting data used to investigate the intra- and inter-host variation in various clinical settings, such as infectious mononucleosis and immunosuppression with chronic EBV infection. Additionally, the number of available whole genomes from East Asia is expanded by eleven (unique) novel genomes from primary infection from a NPC-non-endemic area. These sequences were used for a comparative analysis between NPC- and non-NPC-derived EBV genomes and a number of sites were determined differentiating these two groups.

Finally, comparative genomic analyses of world-wide EBV strain diversity were performed using genome sequences generated here in conjunction with a large number of publicly available EBV genome sequences. The comprehensive data sets generated, which included measures of diversity, selection, and linkage, were used to identify potential targets of T cell immunity. In addition, the population structure of EBV was analysed to better understand the forces that have shaped the evolution of EBV.

Acknowledgements

First and foremost, I would like to thank my primary supervisor, Professor Judy Breuer, for her advice and guidance throughout my PhD, her compassion, and the many opportunities she provided to support my scientific career.

I would like to thank Professor Benny Chain, my secondary supervisor, for always being available to offer advice on the project.

A very large thank you also goes to Dr Dan Depledge, who was my *de facto* supervisor regarding all practical issues in the lab. I am very grateful for the good advice and help I have received from him, and even more so for the numerous discussions of scientific nature and beyond.

More generally, I would like to acknowledge the help received from other scientists on this project: Dr Florent Lassalle, Professor François Balloux, Dr Jamie Heather, Dr Sofia Morfopoulou & many more besides.

I would also like to say a big thank to my friends for their support throughout. Zu guter Letzt möchte ich ein großes Dankeschön meinen Eltern aussprechen. Ohne eure uneingeschränkte Unterstützung existiere diese Arbeit heute nicht.

Contents

List of Figures

List of Tables

List of Abbreviations

(c)HL	(classical) Hodgkin's lymphoma
(q)PCR	(quantitative) PCR
(s)LCL	(spontaneous) lymphoblastoid cell line
BART	BamHI A rightward transcript
BCR	B cell receptor
BL	Burkitt's lymphoma
CDS	Coding DNA sequence
CMV	Cytomegalovirus
CTL	Cytotoxic T cell
CVC	Capsid vertex component
DNA	Deoxyribonucleic acid
EBER	Epstein–Barr virus-encoded small RNAs
EBNA	EBV nuclear antigen
EBV	Epstein Barr virus
EBVaGC	EBV-associated gastric carcinoma
GC	Germinal centre
gp	Glycoprotein
HLA	Human leukocyte antigen
HSCT	Hematopoietic stem cell transplantation
HSV-1	Herpes simplex virus 1
IC	Immunocompromised sample set
IEDB	Immune epitope database
IG	Immunogenic
IM	Infectious mononucleosis
LCV	Lymphocryptoviruses
LD	Linkage disequilibrium
LMP	Latent membrane protein
MHC	Major histocompatibility complex
miRNA	micro RNA
ML	Maximum likelihood
MS	Multiple sclerosis
NGS	Next generation sequencing
NHP	Non-human primate
NIG	Non-immunogenic
NJ	Neighbor joining
NPC	Nasopharyngeal carcinoma

ORF	Open reading frame
OTR	On-target read
PCA	Principal component analysis
PHI	Pairwise homoplasy index
PSS	Positively selected site
PTLD	Posttransplant lymphoproliferative disorder
RNA	Ribonucleic acid
SHM	Somatic hypermutation
SNP	Single nucleotide polymorphism
SOT	Solid organ transplantation
TB	Paired tumour and blood sample set
VCA	Vertex capsid antigen
VL	Viral load
WGA	Whole genome amplification
WGS	Whole genome sequencing

Chapter 1

Introduction

1.1 A brief evolutionary history of Lymphocryptoviruses

Epstein-Barr virus (EBV) belongs to the family of Herpesviridae (Davison et al., 2009) and is spread world wide in the vast majority of the human population. It further belongs to the genus of Lymphocryptoviruses (LCV) of the sub-family of γ-herpesviruses (the other sub-families being α- and β-herpesviruses).

The natural hosts of γ-herpesviruses are mammals. The separation into the three herpesvirus sub-families occurred probably 400 million years ago (MYA), which was long before the mammalian radiation 60-80 MYA. But the divergence into the existing species was estimated to have happened over the last 60 MY, showing the long period of coevolution between herpesviruses and their mammalian hosts (McGeoch et al., 1995).

Within human and non-human primates (NHP), New World and Old World LCVs form distinct sister clades (McGeoch, Gatherer, and Dolan, 2005). The New World LCV phylogeny based on viral DNA polymerase sequences generally reflects the phylogeny of their respective hosts; in Old World LCVs however, the branching pattern is not well resolved and even with extended sequence data sets, the phylogeny remained incomplete due to multifurcations and low statistical support of nodes. This, however, also suggests that these viruses evolve more slowly as sequence variability is lower and could be indicative of a more complex evolutionary history than just a synchronous host and virus coevolution (Ehlers et al., 2010; Lacoste et al., 2010).

Human and NHP LCVs share a similar biological lifestyle. They establish a lifelong latent infection, in which the virus is maintained as an episomal DNA molecule with only a very reduced protein expression profile. Replication of the virus and production of virions occurs during a lytic cycle. Nuclear antigens responsible for the establishment of latency are functionally conserved, but show a high degree of sequence divergence. In consequence, simian nuclear antigens do not cross-react with human anti-EBV sera (Gerber et al., 1977). Lytic genes, however, are more conserved across EBV and NHP LCVs, and cross-reactivity of human and Old World NHP antibodies has been shown (Cho et al., 1999).

Interestingly, comparison of genome content and organisation between a New World LCV genome (derived from CalHV3, a common marmoset-derived cell line) and Old World LCV genomes including EBV highlighted eleven genes that are specific to Old World LCVs, i.e. they were most likely acquired after separation of New World and Old

World primates (Rivailler, Cho, and Wang, 2002). These genes are not essential for transformation or replication, but rather facilitate host invasion, explaining the higher prevalence of LCV infection in Old World primates compared to New world primates. The highest degree of similarity within the Old World LCV group as well between EBV and New World NHP LCV has been found in the Epstein Barr nuclear antigen 1 (*EBNA1*) homologue, possibly due to its essential role in maintenance of the viral genome (Yates et al., 1996; Blake et al., 1999; Lacoste et al., 2010).

1.2 Epstein-Barr virus

EBV was discovered in 1964 by Sir Anthony Epstein and Yvonne Barr in a Burkitt's lymphoma (BL) cell line via electron micrographs, and was the first virus being associated with the development of human tumours (Epstein, Achong, and Barr, 1964). Since its discovery EBV has been associated with a variety of other cancerous diseases as well as infectious mononucleosis (IM) (Henle, Henle, and Horwitz, 1974). It has been estimated that 1.8 % of cancer related deaths are due to EBV-associated malignancies (Khan et al., 2014).

Like all herpesviruses, it consists of a large double stranded DNA molecule, which is enclosed by a capsid, tegument, and envelope (Davison et al., 2009). EBV is extremely successful in infecting humans. It is the most common viral infection as more than 95 % of the world population are persistent carriers for life (Young, Yap, and Murray, 2016).

1.3 The life cycle of EBV

1.3.1 Virus entry

EBV is transmitted via saliva into which infectious virus particles are shed from already infected hosts. Infection occurs usually very early in life. In developing countries, the peak of infection as measured by EBV seropositivity is usually seen in the early years of life (3-4 years), while there are often two peaks in developed countries: one peak in early age (below 5 years) and one during adolescence (after 10 years) (Hjalgrim, Friborg, and Melbye, 2007).

The primary infection targets are B cells and epithelial cells, although under some circumstances EBV can also infect T cells, Natural Killer (NK) cells and others (Isobe, 2004). The virus usually enters the body through the oropharynx and encounters B cells in the tonsils. Whether or not replication in epithelial cells plays also a role during infection is still not clear, but recent evidence points into this direction. However, replication in epithelial cells is likely an early event and in IM patients, the only disease associated with primary infection, EBV cannot be found in epithelial cells of the oropharynx.

Host cells are entered using a mechanism involving at least five envelope glycoproteins (gp). The core glycoproteins shared and conserved among all herpesviruses are

gH/gL and gB, whereas the tropism is determined by gp350 and gp42. The primary targets of infection are B cells through interaction between viral gp350 and gp42 with host CD21 and HLA class II on the cell surface (Nemerow et al., 1987; Spriggs et al., 1996), which leads to fusion with the B cell membrane. Epithelial fusion, on the other hand, is mediated through interaction of viral gH and host α-v integrins (Chesnokova and Hutt-Fletcher, 2011) as well as viral BMRF2 protein with host β-1 integrins (Xiao et al., 2008).

Interestingly, the viral tropism seems to be dependent on virion origin. Virions produced in B cells have fewer gp42 molecules than virions produced in epithelial cells and vice versa. This is due to HLA class II processing in B cells, which renders the interacting complex more prone to degradation and HLA class II presentation (Borza and Hutt-Fletcher, 2002). Epithelial cells, however, do not possess HLA class II on their surface and virions produced here are therefore gp42-enriched. As a consequence, virions originating from epithelial cells are much more infectious for B cells than virions originating from B cells (Jiang, Scott, and Hutt-Fletcher, 2006). Likewise, B cell-produced virions are more infectious for epithelial cells than those produced in epithelial cells. This is because the presence of gp42 in the glycoprotein complex hinders the interaction with integrins (Hutt-Fletcher, 2016). These observations of a tropism switch have led to the hypothesis that EBV naturally alternates replications in the two cell types.

1.3.2 EBV latency

The virus' life cycle has two stages: latency and lytic replication.

During latency, the virus remains within host cells as a circular extrachromosomal DNA molecule and displays a very restricted gene expression profile in order to avoid immune detection. Latency is established in memory B cells (Babcock et al., 1998), but there is still uncertainty about the route of how this happens. There are two, not mutually exclusive, models of persistence: The germinal centre (GC) model and the direct infection model.

The germinal centre model According to this model, EBV^+ memory B cells are the result of normal B cell differentiation (figure 1.1) (Thorley-Lawson and Gross, 2004; Babcock, Hochberg, and Thorley-Lawson, 2000). B cell differentiation usually occurs after naïve B cells become exposed to an antigen. They enter the germinal centre, a structure within the lymph nodes, where they undergo the germinal centre reaction and proliferate. Their immunoglobulin genes undergo class switching and somatic hypermutation (SHM). The positive selection of antigen-specific B cells is further assisted by T cells through CD40 receptor and antigen activation of the B cell receptor (BCR). B cells exit the germinal centre reaction differentiated to either (antibody producing) plasma cells or memory B cells.

In the case of EBV infection, however, naïve B cells are infected with the virus in the lymphoid tissue in the oropharynx. Instead of being antigen-exposed, the virus expresses a program called *latency III* (or growth program), which drives the proliferation and expansion of infected B cells (Thorley-Lawson and Gross, 2004). This program is

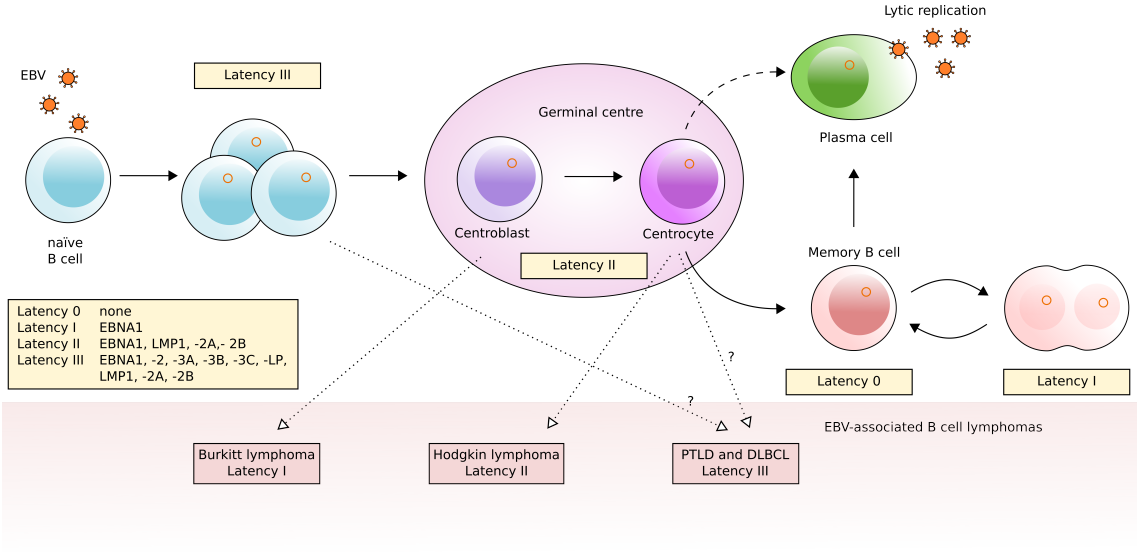


FIGURE 1.1: Upper panel: Germinal centre model of virus persistence. EBV infects naive B cells, which are driven to proliferation by the latency III program. The lymphoblasts enter the GC, where they mature and receive survival signals by the latency II program. EBV infected immortalised B memory cells exit the GC and circulate in the blood of the host without any viral gene expression except during cell division. EBV-infected memory B cells can transform into plasma cells upon antigen recognition, which also results in lytic virus replication. It might also be possible that EBV-infected plasma cells directly exit from the GC.
Lower panel: The origin of EBV-associated lymphomas. The exact stages are not really known and patterns of latency between progenitor and tumour cell cannot be expected to correspond. Figure modified from Young, Yap, and Murray, 2016

characterised by the expression of nine latency genes: six nuclear antigens (EBNA1, -2, -3A, -3B, -3C, and -LP), three latent membrane proteins (LMP1, -2A, and -2B) as well as EBV-encoded small RNA (EBER) and non-transcribed BART (*Bam*HI-A region rightward transcript) miRNAs.

These EBV$^+$ cells enter the germinal centre reaction. Here, a program called *latency II* (or default program) is expressed, which consists only of EBNA1, LMP1 and LMP2. The CD40 and BCR signalling is mimicked through LMP1 and LMP2 which are functional homologues of the respective receptors, thereby providing the necessary survival signal in the GC (Gires et al., 1997; Caldwell et al., 1998).

The transformed EBV-infected memory B cells finally exit the GC and circulate in the blood stream of the host. At this stage, in order to avoid immune detection, EBV downregulates viral gene expression (*latency 0*), but expresses EBNA1 (*latency I*) during cell division.

The direct infection model The germinal centre model is inconsistent with observations of B cells isolated from GCs of IM patients, which do not seem to undergo SHM, which would be expected if they partook in the GC reaction. Moreover, they exhibited an unusual expression pattern by expressing EBNA2 but not LMP1, thereby differing from

EBV^+ GC B cell-derived tumours (BL, HL, and PTLD). It has therefore been proposed that EBV in IM directly infects GC and/or memory B cells which expand without SHM (Kurth et al., 2003).

1.3.3 Latency proteins

EBNA1 is the only protein consistently expressed in all latency programs (in latency 0 only during cell division). It has several vital functions: initiating replication of the viral genome, mitotic segregation and acting as a transcription factor for other latency proteins. It is a homo-dimeric protein that initiates EBV genome replication by binding to the EBV latent origin of replication (oriP). Furthermore, it assures the equal partitioning of viral genomes to the host daughter cells. The viral DNA EBNA1-binding site is found in the family of repeats (FR) region of the EBV genome, where it functions as a transcriptional transactivator and enhances the transcription of other latent genes (Gahn and Sugden, 1995). It also binds to cellular promotors thereby regulating the expression of host genes (Canaan et al., 2009). It is further involved in p53 degradation and oncogenesis (Fries, Miller, and Raab-Traub, 1996).

EBNA-LP is essential for B cell transformation and interacts with EBNA2 and other transcription factors to activate viral and host gene transcription. EBNA2 acts as the main transcription factor essential for B cell transformation by upregulating both groups of viral latent genes (EBNAs and LMPs) as well as host genes including the B cell activation molecules CD21 and CD23, as well as *MYC* (Pan et al., 2009; Kaiser et al., 1999). Instead of binding DNA directly, it interacts with other host DNA-binding proteins like RBP-Jκ (Tzellos et al., 2014).

A coactivator of EBNA2 is EBNA3A, which also interacts with EBNA3C and inhibits RBP-Jκ recruitment, downregulates *MYC*, leads to arrest of host cells G1 phase, and blocks EBNA2 activation effects (Cooper et al., 2003). It is essential for B cell transformation. Similarly, EBNA2 is coactivated by EBNA3B. However, it has been shown not to be essential for B cell transformation (Tomkinson, Robertson, and Kieff, 1993). Instead, it was suggested to act as a tumour supressor (White et al., 2012). The third and most essential EBNA2 coactivator is EBNA3C. Together they upregulate the chemokine CXCL12 and its receptor CXCR4 (Zhao et al., 2011), which have been shown to be essential for lymphoblastoid cell lines (LCL) proliferation and cell survival in a mouse model (Piovan et al., 2005). It is also involved in a wide array of regulatory functions influencing B cell growth such as the repression of the tumour suppressor $p16^{INK4a}$ in cooperation with EBNA3A (Skalska et al., 2013). Additionally, EBNA3C can block apoptosis by inhibiting the EBV-infection-mediated DNA damage response (Saha et al., 2012).

The *LMP1* gene encodes for a CD40 homolog (Gires et al., 1997). It activates NF-κB, JNK and p38 pathways and is therefore a major oncogene (Young, Arrand, and Murray, 2007; Izumi and Kieff, 1997).

LMP2A promotes cell growth. By activating ERK/MAPK pathway constitutively, it mimics BCR signalling (Caldwell et al., 1998; Anderson and Longnecker, 2008). Simultaneously, it inhibits antigen-dependent BCR signalling. It is involved in the inhibition of

apoptosis and essential in rescuing germinal center B cells which lack a functional BCR. Additionally, it promotes epithelial cell spreading. LMP2B, a splice variant, negatively regulates LMP2A (Wasil et al., 2013).

1.3.4 Lytic cycle

During the lytic cycle, new EBV virions are produced and it is therefore a necessity for host to host transmission. Infectious virus is commonly found in the saliva of immunocompetent, asymptomatic hosts (Ling et al., 2003; Hadinoto et al., 2009), suggesting that there are sites of lytic replication within or close to the oral cavity.

Lytic cycle is usually only entered after reactivation from latency, but could also be initiated upon primary infection, e.g. in epithelial cells. However, due to the long incubation period for IM, primary lytic replication has not been observed (Kenney, 2007). The lytic cycle can be initiated by BCR stimulation of B cells and the consequent differentiation into plasma cells (Laichalk and Thorley-Lawson, 2005).

The transcription hierarchy can be divided into immediate-early (IE), early (E) and late (L) genes. The two IE genes encoding transactivators are *BZLF1* and *BRLF1* which encode the major transactivators ZEBRA/Zta and Rta, respectively. Both transcription factors are responsible for the switch from latent to lytic phase. They allow the subsequent expression of E gene products, which include viral replication proteins such as the viral *Pol*, a heterodimer comprised of two subunits encoded by *BALF5* (catalytic) and *BMRF1* (accessory). In total there are six core viral replication proteins: BALF5, BMRF1, BALF2 (binds single stranded DNA), BBLF4 (helicase), BSLF1 (primase), and BBLF2/3 (primase-associated protein) (Fixman, Hayward, and Hayward, 1992). Other E gene products are responsible for deoxynucleotide metabolism, whereas many L genes encode structural proteins and glycoproteins, such as *BcLF1* (major capsid protein), *BLLF1* (gp350), *BILF2* (glycoprotein) and others (Kenney, 2007).

In some cases, full EBV genome integration into the host genome has been reported in cell lines (Matsuo et al., 1984). Despite evidence for expression of latency proteins (Luo et al., 2004), reactivation does not seem to occur from these genomes. In nasopharyngeal carcinoma (NPC) biopsies, integrated EBV genomes could also be found in some cases, indicating it can occur *in vivo* as well as *in vitro*. Moreover, coexistence of integrated EBV genomes and episomes has been found in B-derived cell lines (Anvret, Karlsson, and Bjursell, 1984; Delecluse et al., 1993).

1.4 The anti-EBV immune response

The establishment of latency results in a life long infection in a peripheral pool of recirculating memory B cells. Upon re-entering the tonsils, memory B cells either change their expression program or enter lytic replication and produce new infectious viral particles, which can infect new naïve B cells or be shed into the saliva (Thorley-Lawson, 2001).

In asymptomatic, immunocompetent carriers, reactivation and expression of viral proteins leads to a rapid **adaptive immune response**, which controls but does not eliminate the infection. The main effectors of this response are T cells. More than 50 epitopes for both HLA class I and II have been described (Hislop et al., 2007). $CD4^+$ and $CD8^+$ T cells differ in their EBV immune targets and there is a hierarchy in the strength of immune responses these targets elicit. The strongest cytotoxic T cell responses have been observed against epitopes derived from proteins of the EBNA3 family as well as IE and E lytic cycle proteins (Mautner and Bornkamm, 2012). In contrast the $CD4^+$ T cell response has been found to be particularly strong against lytic cycle proteins, especially structural antigens, followed by autoantigens and then latent proteins (Mautner and Bornkamm, 2012).

There is also an antibody-mediated response that is established during primary infection, which partly persists for life in form of immunologbolulin (Ig) G antibodies directed against the viral capsid antigen and EBNA1 (see subsection 1.5.1, kinetics shown in figure 1.2) (Henle et al., 1987; Hinderer et al., 1999). These responses, in addition to IgM against viral capsid antigen (VCA), appear at specific phases of acute primary infection and are therefore of diagnostic value. Further targets of the humoral response are early antigen diffuse (EA-D) and gp350 (Panikkar et al., 2015).

Additionally, the **innate immune system** acts against viral infections by recognition of viral particles and nucleic acids through Toll-like receptors (TLRs) which activate an interferon (IFN) response. For example, EBERs, which are released from EBV-infected cells and are present in sera of patients with a variety of EBV-related pathologies (including IM, chronic active EBV infection, and EBV-associated hemophagocytic lymphohistocytosis), induce TLR3 signaling (Iwakiri et al., 2009). Other TLRs which have been shown to activate an innate immune response against EBV are TLR9 and -7 (Quan et al., 2010), TLR2 (Ariza et al., 2009) and TLR8 (Farina et al., 2017).

Conversely, EBV has evolved mechanism to evade both adaptive and innate immune responses. For example, BPLF1, the large tegument protein, assists in innate immune evasion by interfering with TLR signaling (van Gent et al., 2014); the late gene *BCRF1* encodes a interleuking (IL)-10 homologue with 84 % sequence identity to human IL-10 on the amino acid level. IL-10 is a immunomodulatory cytokine that acts as a suppressor of T cell proliferation and cytokine production and inhibits interferon (IFN)-γ production (Mosser and Zhang, 2008); interestingly, it has even been suggested that anti-gp350 antibodies enhance infection of epithelial cells, thereby providing an alternative reservoir for EBV during a T cell response against EBV-infected B cells (Turk et al., 2006).

1.5 EBV-associated diseases

EBV is for the most part non-pathogenic and has evolved complex strategies to establish a life-long infection but to remain undetected. However, due to the nature of the transformations it induces in B cells, it is associated with a variety of diseases, most of them cancerous.

1.5.1 Infectious mononucleosis

Infectious mononucleosis (IM) is the only disease caused by EBV associated with primary infection. It was first described in 1920 (Sprunt and Evans, 1920) and is characterised by pharyngitis, cervical lymph node enlargement, fatigue, and fever. It further shows atypical large cells in the blood that have been later identified as $CD8^+$ T cells, which are responding to EBV-infected B cells (Callan et al., 1998).

Children become EBV^+ at a younger age in developing countries than in developed countries, where IM is most often observed in adolescents and young adults and where it is thought to be acquired due to kissing (hence IM's nick name 'kissing disease') (Balfour, Dunmire, and Hogquist, 2015). One reason why IM is less frequently diagnosed in younger children is that intimate kissing probably transmits large amount of infectious virus, whereas younger children most likely become infected through parents and siblings who transmit smaller amounts of oral secretions and therefore virus. However, IM in children younger than 12 years has been observed and is not uncommon (Horwitz et al., 1981), but heterophile antibody tests are often unreliable in this patient group (Balfour, Dunmire, and Hogquist, 2015). It has also been proposed that IM arises due to cross-reactive influenza-specific $CD8^+$ T cell responses, which are more likely to have been acquired in high numbers at later age (Clute et al., 2005), but this hypothesis is still contentious.

EBV can also be transmitted via blood, which suggests that the virus found in (memory) B cells is or can become infectious (Alfieri et al., 1996). Additionally, transmission via transplantation of stem cells or solid organs can lead to life-threatening complications, in particular in EBV^- recipients (see Post-transplant lymphoproliferative disorders).

The incubation period of IM is approximately six weeks (Hoagland, 1955). Due to the long incubation period and the subtle onset of disease, little is known about the early virus-host interactions of primary infection. There are three phases of infection: acute infection (0-3 weeks after onset), sub-acute phase (4 weeks-3 months) and convalescence (4-6 months). During these stages, a discrete set of antibodies is being produced. IgM antibodies targeting the VCA are first produced by around 75 % of patients (Hinderer et al., 1999; Balfour et al., 2013) and can be detected during acute and sub-acute phase but not during convalescence. IgG antibodies against VCA (produced as early as two weeks after onset) (Hinderer et al., 1999) and IgG against EBNA1 (only during convalescence) are produced by nearly all patients and persist for life (figure 1.2) (Henle et al., 1987). The late onset of an antibody response against EBNA1 correlates with a delayed CD4 T cell response against EBNA1 (Long et al., 2013).

EBV viral loads in blood and oral compartments increase sharply during acute infection and subsequently decrease over time, though specific kinetics can vary between patients. However, the virus is cleared faster from blood than from the oral compartment and oral shedding of infectious virus can continue for several months (and recur upon reactivation).

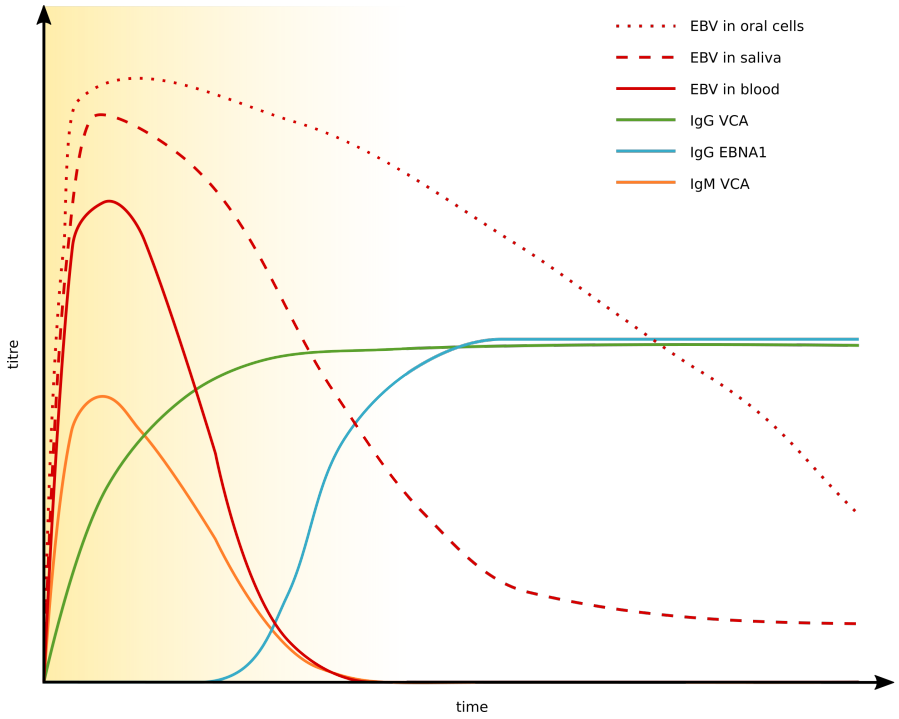


FIGURE 1.2: EBV kinetics of antibody titres and viral loads during IM. The background colour relates to the phases of IM infection (yellow: acute phase and sub-acute phase, white: convalescence). IgM antibodies targeting VCA are produced by the majority of patients during acute and sub-acute phase but not during convalescence. Shortly after onset of disease, IgG against VCA are produced by virtually all patients and remain high for the rest of the life. EBNA1 IgG antibodies are being produced during convalescence and also persist for life. Viral load increases sharply in all compartments and is generally higher in oral compartments where it also decreases at a slower rate during convalescence compared to blood. Graph modified from Odumade, Hogquist, and Balfour, 2011.

1.5.2 Burkitt's lymphoma

Burkitt's lymphoma (BL) was first described by Denis Burkitt in 1958 occuring in children in equatorial Africa (Burkitt, 1958) and was the source from which EBV was first isolated (Epstein, Achong, and Barr, 1964). It is a rapidly proliferating tumour usually of B cell origin. There are three forms of BL: the endemic, high incidence form (also called the 'African' variant and the most common childhood malignancy in equatorial Africa), the sporadic form (occurring throughout the world) and the HIV-associated form (Vockerodt et al., 2015). Common to all forms is the overexpression of *MYC*, a transcription factor involved in regulation of many growth and proliferation-related genes (*Uniprot Database* accessed last April 2017). The overexpression occurs due to a reciprocal translocation of the *MYC* oncogene and one of the Ig light or heavy chain loci, which brings *MYC* under the transcriptional control of an Ig locus (Klein, 1983) and drives the proliferation of the lymphoblasts at a high rate. This translocation is generally seen as the essential step in BL formation while EBV plays a supportive role by providing rescue signals for aberrant BL cells.

The EBV association varies between BL types. Endemic BL is almost always EBV$^+$,

whereas incidence rates in sporadic ($\approx$ 10-15 %) and HIV-associated forms ($\approx$ 40 %) are lower (Vockerodt et al., 2015). Moreover, sporadic and endemic forms differ in breakpoint location within the Ig locus (VDJ/VJ regions in endemic BL, switch region in sporadic BL). One the one hand, this suggests that BL cells in endemic and sporadic cases may have acquired their *MYC* translocation during different phases of B cell differentiation (pre-GC and GC, respectively). On the other hand, VDJ/VJ breakpoints could occur in the GC as a result of abnormal activation-induced deaminase (AID) activity, which is responsible for inducing SHM in immunoglobulins (Goossens, Klein, and Küppers, 1998). AID and SHM have been shown to be induced by LMP1 expression (Epeldegui et al., 2007), and in fact, SHM occurs more frequently in endemic and HIV-associated BL than in sporadic BL (Bellan et al., 2005). However, AID is in turn also inhibited by EBNA2 (Tobollik et al., 2006). Moreover, BL usually display a latency I profile in which only EBNA1 is expressed. In consequence, the role of SHM induction through EBV *in vivo* is not completely clear.

However, EBV provides another key component for MYC-driven oncogenesis. Besides inducing cell proliferation, MYC overexpression also induces apoptosis (Milner et al., 1993), but a number of EBV proteins have anti-apoptotic functions, among them EBNA1 which is critical for EBV$^+$ BL (Kennedy, Komano, and Sugden, 2003).

Another factor playing into the development of endemic BL are coinfections. The high-incidence area of endemic BL overlaps with areas of holoendemic *Plasmodium falciparum* infection, which causes malaria. Two mechanisms could explain this observation: First, malaria can lead to loss of EBV specific T cell immunity due to functional exhaustion of T cells (Wykes et al., 2014). Second, malaria antigens themselves activate the immune system and thereby drive polyclonal B cell expansions as well as lytic EBV reactivation which increases the pool of EBV$^+$ B cells. In turn, this increases the chance of acquiring a *MYC* translocation (Chêne et al., 2007).

1.5.3 Hodgkin's lymphoma

This malignancy was first described by Thomas Hodgkin in 1832 and its cause is still unknown, even though the involvement of an infectious agent had been discussed early on. The malignant cell type characteristic of Hodgkin's lymphoma (HL) is Hodgkin/Reed-Sternberg (HRS) cells. The vast majority of them are derived from GC B cells, i.e. they have undergone SHM but do not possess a functional BCR due to crippling mutations in the Ig genes (Küppers et al., 1994). These tumour cells are also extremely rare, making up only 1 % of the tumour mass, as they are surrounded by a massive inflammatory infiltrate made up of T cells, histiocytes, eosinophilic granulocytes and plasma cells (Bräuninger et al., 2006).

EBV is the most likely candidate of infectious agents to be responsible for the transformative events leading to tumourigenesis. In almost 40 % of cases of classical Hodgkin's lymphoma (cHL), HRS cells are EBV$^+$ and express the latency II program. The virus might provide anti-apoptotic signals for these cells, which under normal conditions would have been selected against in the GC. EBV has been shown to be able to rescue GC B cells

with "crippled" BCRs as well as B cells not expressing any BCR *in vitro* through BCR-like signaling via LMP2A (Bechtel et al., 2005; Chaganti et al., 2005; Mancao et al., 2005). Moreover, all cHL cases with crippling mutations that prevent BCR expression were found to be EBV^+ (Bräuninger et al., 2006). Additionally, a number of pathways in HRS cells are aberrantly activated, such as NF-κB, JAK/STAT and PI3K/AKT, all of which can be activated through LMP1.

The involvement of EBV in the development of cHL is further supported by the finding that there is an increased risk of developing EBV^+ HL (but not EBV^-) after infectious mononucleosis (IM), in particular in young adults (Hjalgrim et al., 2003).

1.5.4 Nasopharyngeal carcinoma

Nasopharyngeal carcinoma (NPC) is one of the few carcinomas – malignancies of epithelial origin – associated with EBV. Globally it is a rather rare tumour, but prevalence rates vary geographically. Areas with particularly high prevalence include North Africa, Southern China, South East Asia, as well as the Inuit people in Alaska (Wei and Sham, 2005). Here, the incidence rate is 20-30 times higher in men and 8-15 times higher in women than in Europe and the rest of America (WHO, 2014). Interestingly, the incidence rate of NPC in Chinese people who have immigrated to North America remains high, but is lower in Chinese people born in North America (Wei and Sham, 2005; Dickson and Flores, 1985), even though it is still elevated compared to local Caucasians (Buell, 1974; Bei et al., 2016). It has been proposed that carcinogenesis is dependent on several factors: host genetics (as highlighted by the distribution among populations), EBV infection and genetics, and environmental factors such as smoking and the consumption of salted fish and preserved/cured meat (Chen et al., 1990; Jia et al., 2010).

NPC is classified into two categories based on the tumour's microscopic appearance: keratinising squamous cell carcinoma (type I) and non-keratinising squamous cell carcinoma (types II and III, which are differentiated and undifferentiated, respectively) (Young and Dawson, 2014). Type I is relatively rare in Southern China, whereas types II and III are the NPC forms in which EBV is consistently found and which are of particular interest in endemic NPC regions (Wei et al., 2011).

The typical latency program expressed in NPC tumours is latency II (i.e. LMP2A/B, LMP1, EBNA1, EBER1 and -2, as well as BART miRNAs). LMP1 expression is variable in NPC, with approximately 20-40 % of tumours expressing it at the protein level (Young and Dawson, 2014). It acts as a major oncogene and has been shown to induce hyperproliferation, drive the production of proinflammatory cytokines, enhance cell motility, and provide anti-apoptotic function in epithelial cells (Dawson, Port, and Young, 2012). The expression pattern of LMP2B is similar. In contrast to that, LMP2A is consistently detected. LMP2A has been shown to be essential for epithelial cell outgrowth *in vitro* (Scholle, Bendt, and Raab-Traub, 2000) and enhance motility and cell adhesion (Allen, Young, and Dawson, 2005; Lu et al., 2006). Moreover, it can induce epithelial-to-mesenchymal transition (EMT), a process connected with the acquisition of stem cell-like properties (Kong et al., 2010).

EBV infection is thought to play a supportive role in NPC pathogenesis and not to be the initiating event, as healthy individuals at high risk of NPC as well as IM patients lack evidence of EBV^+ epithelial cells in the oropharynx (Young and Dawson, 2014). Other important factors are genetic and epigenetic changes found in NPC which possibly precede and are required for infection. According to current model of NPC tumourigenesis, environmental factors lead to the loss of heterozygosity on chromosomes 3p and 9p. As a consequence, low-grade, preinvasive lesions form, which are susceptible to stable EBV infection, in particular after additional genetic changes (such as overexpression cyclin D1 and the creation of a undifferentiated cellular environment) (Young and Dawson, 2014). However, EBV infection likely ocurs before clonal expansion of tumour cells, as it has been found to be monoclonal in NPC tumours (Raab-Traub and Flynn, 1986).

1.5.5 Post-transplant lymphoproliferative disorders

EBV infection is controlled mainly through T cell immunity, but NK cells and antibody responses also play a role. Immunosuppression is the standard treatment for recipients of solid organ and haemopoeitic stem cells. As a consequence, T cell responses are weakened which leads to loss of control over EBV replication. In serious cases, this can result in post-transplant lymphoproliferative disorder (PTLD), a term for a heterogeneous collection of diseases ranging from B cell proliferations to lymphoma (Gulley and Tang, 2010).

According to the WHO, there are four histopathologic subtypes of EBV: early lesions, polymorphic PTLD, monomorphic PTLD, and cHL-type PTLD. Early lesions, compared to the other types, are usually polyclonal, but if untreated, a single clone can potentially outgrow to one of the other subtypes (Swerdlow et al., 2008). Polymorphic PTLD consists of lymphocytes and lymphoblasts of varying size, containing B cells as well as $CD4^+$ and $CD8^+$ T cells, while monomorphic PTLD resembles conventional lymphoma, such as DLBCL and immunoblastic lymphoma, and is usually of B cell origin (Gulley and Tang, 2010) and only rarely of T or NK cell origin (Swerdlow, 2007). Classical HL-type PTLD is a rarer type of PTLD (Pitman et al., 2006).

PTLD usually develops during the first year after transplantation: For haematopoeitic stem cell transplants (HSCT), the median onset is two months, for solid organ transplants (SOT) it is six months (Gulley and Tang, 2010). Nearly all PTLD tumours are EBV^+ (60-80 % of cases). While EBV infection is extremely prevalent in the patient population, only a fraction of transplant recipients will develop PTLD. Risk factors are age, EBV seronegativity at the time of transplantation, coinfection with other viruses such as cytomegalovirus (CMV), active EBV disease during transplantation, the intensity of immunosuppression, HLA type and HLA mismatch as well as having a combination of several risk factors (Gulley and Tang, 2010). PTLD occurs more frequently in children, most likely due to the higher rate of primary infections. Additionally, the organ type of the transplant influences the incidence of PTLD (table 1.1).

PTLD tumours usually express the latency III program (i.e. all latency proteins including EBERs and BART miRNAs). Acquired mutations on the host side, however, are not well characterised.

Organ	Incidence (%) in children
Kidney	1-10
Marrow and stem cell	13
Liver	4-15
Heart or lung	6-20
Intestinal	12

TABLE 1.1: Incidence of PTLD by organ type in children (Gulley and Tang, 2010).

To avoid lymphoma formation, reduction of immunosuppression is the first treatment option, so that the host can generate their own adaptive immune responses against EBV and retrieve control over B cell proliferation. This is especially true for early lesions, while it is often but not always sufficient for mono- and polymorphic PTLD. Other treatment options include Rituximab, an anti-CD20 antibody, antivirals such as acyclovir and ganciclovir, and adoptive immunotherapy with EBV specific cytotoxic T cells (Green and Michaels, 2013). Levels of EBV in the blood or plasma is the primary marker for deciding upon treatment for PTLD, but this is not necessarily predictive, as some patients with high viral loads do not develop PTLD (Gulley and Tang, 2010; Kerkar et al., 2010).

1.5.6 Other associated malignancies

Gastric carcinoma is the third leading cause of death related to cancer (Ferlay et al., 2015). Around 10 % of gastric carcinoma tumour are EBV^+ and are phenotypically as well as clincally very distinct from EBV^- gastric carcinoma. Characteristics include highly methylated CpG islands, wild-type p53 expression, loss of p16 tumour surpressor protein, and a distinct pattern of allelic loss (Lee et al., 2004; Schneider et al., 2000; van Rees et al., 2002). Additionally, EBV-associated gastric carcinoma (EBVaGC) displays differences in prevalence between sex, age and ethnicity: it is more common in Caucasian and Hispanics than in Asians, more common in young than in the elderly, and more common in men than women (Lee et al., 2009). Similar to the other epithelial EBV-associated cancer, NPC, EBVaGC tumours express EBNA1, LMP2A, but also BARF1 protein, the EBERs, and BART miRNAs (zur Hausen et al., 2000; Imai et al., 1994). EBV is absent in premalignant lesions, suggesting again that EBV infection is a late event in gastric carcinogenesis (zur Hausen et al., 2004).

All malignancies described so far are associated with latent infection. The only disease caused by lytic EBV infection is **oral hairy leukoplakia**, an AIDS-associated epithelial hyperplasia on the lateral tongue (Hutt-Fletcher, 2016; Greenspan et al., 1985). It can be treated with antiviral drugs which inhibit lytic replication such as acyclovir.

Multiple sclerosis (MS) is a EBV-associated, but non-cancerous disease. It is a chronic demyelinating disease of the central nervous system (CNS), which causes severe progressive disability especially in younger people. It affects more than 2.5 million people worldwide and is more common in females than males (Pender and Burrows, 2014). While usually seen as an autoimmune disease, more and more evidence suggests that EBV plays an important role in pathogenesis. A meta-analysis showed that 100 % of MS patients are EBV^+ with two independent methods of EBV detection (Pakpoor et al., 2013). Second, a history of IM increases the risk of MS (Thacker, Mirzaei, and Ascherio, 2006). Most studies suggest now that EBV infection is a prerequisite of developing MS, but not sufficient on its own as only a small subset of EBV^+ carriers develop the disease (Pender and Burrows, 2014). There are four mechanistic hypotheses about the possible contribution of EBV to MS:

1. The cross-reactivity hypothesis suggests that EBV-specific T cells cross-react with CNS antigens (Wucherpfennig and Strominger, 1995). However, a substantial body of evidence disproved cross-reactivity to be the main cause of MS development, even if it might contribute to disease progression (Pender and Burrows, 2014).

2. The bystander-damage hypothesis proposes that the damage caused by the immune system in the CNS is primarily a bystander-effect of the immune response against EBV antigens and not due to an autoimmune reaction (Serafini et al., 2007).

3. The αB-crystallin or 'mistaken self' hypothesis builds on the observation that infectious agents such as EBV can cause the expression of αB-crystallin, a small heat-shock protein and known immunodominant antigen in MS patients, in lymphocytes. The immune system mistakes this protein as a microbial antigen and mounts a CD4-mediated immune response. However, αB-crystallin is also produced naturally by oligodendrocytes in the CNS, and the wrongly directed immune response results in their demyelination (van Noort et al., 2000).

4. The EBV-infected autoreactive B cell hypothesis proposes that autoimmune diseases including MS are caused by EBV infection of already autoreactive B cells. They accumulate in their respective target organ due to a genetically determined defect in the CD8 response against EBV^+ B cells, where they produce autoantibodies and provide costimulatory survival signals for autoreactive T cells (Pender, 2003).

1.6 Prevention and treatment of EBV infection

1.6.1 EBV Vaccine development

Vaccines have been successfully developed against a number of herpesviruses. There is a Varicella Zoster virus (VZV) vaccine, first developed in the 1970s in Japan, which prevents the development of chickenpox and reduces the rate of zoster – the two main diseases caused by this virus in humans – but does not prevent infection (Takahashi et

al., 1974). Additionally, there are vaccines for two non-human herpesviruses: gallid herpesvirus 2, which causes Marek's disease in chickens (Churchill, Chubb, and Baxendale, 1969; Witter et al., 1970), and herpesvirus saimiri, a γ-herpesvirus infecting monkeys closely related to human Karposi's sarcoma-associated virus (KSHV) (Ablashi and Easton, 1976).

For EBV, however, vaccine development has been rather slow. Most efforts target gp350, the most abundant surface protein on virions and EBV-infected cells (Johannsen et al., 2004), responsible for the endocytosis of virions during cell entry. However, to date, only one phase II clinical trial has been reported (Sokal et al., 2007). It was based on soluble gp350 and reduced the rate of IM by 78 %, but did not prevent infection.

Alternatively, peptides of EBV proteins could be used to induce T cell immunity. For example, Elliott et al., 2008 administered EBNA3A peptides to seronegative HLA-controlled subjects. While T cell responses were developed by eight of nine subjects, and no adverse events were reported, infection could not be prevented.

Yet another strategy uses virus-like particles, where latency proteins and transactivators like BZLF1 are either inactivated or deleted, and which lacked the TR packaging element normally required for virion DNA packaging (Ruiss et al., 2011). In mice, both neutralising antibodies as well as cellular immune responses could be shown, but manufacturing is arguably difficult especially for usage in humans (Cohen, 2015).

1.6.2 Therapeutic EBV-specific T cell infusions

Administration of EBV-specific $CD8^+$ T cells is an alternative therapeutic option, which has been demonstrated to show activity in different EBV-associated malignancies. The administered cytotoxic T cells (CTLs) need to be HLA-matched to the recipients in order to be effective and avoid rejection. Both, autologous CTLs and HLA-matched EBV-CTLs from EBV-seropositive donors can be used (Nijland et al., 2016). This approach is investigated and results have been promosing in the setting of for example post-transplant care of HSCT as well as SOT (Doubrovina et al., 2012; Icheva et al., 2013; Comoli et al., 2002; Haque et al., 2007), HL and non-HL (Roskrow et al., 1998; Bollard et al., 2014), and NPC (Straathof et al., 2005; Smith and Khanna, 2012).

1.6.3 CRISPR/Cas9 system as a therapeutic strategy

The CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats/CRISPR–associated) system has evolved in bacteria and archaea as a defense mechanism against viruses. It enables the cells to degrade foreign nucleic acids of viruses and mobile genetic elements. Recently, this system has been developed into a genetic engineering tool (Ran et al., 2013). It consists of two components: Cas9, a bacterial endonuclease, and guide RNA (gRNA) which contain guide sequence. By co-expressing Cas9 and gRNAs of interest, it is possible to recruit Cas9 to almost any site of interest in the genome where it cleaves the double stranded DNA. DNA double strand breaks are repaired by the non-homologous end joining pathway, an error-prone DNA repair mechanism of mammalian

cells. Insertions and deletions can therefore easily lead to a frameshift or premature stop codons, which disrupt the ORF.

Wang and Quake, 2014 first applied the CRISPR/Cas9 system in Raji cells as a model for BL. They were able to demonstrate the clearance of EBV in a subpopulation of cells. They used seven guide RNAs, thereby effectively targeting three different categories of sequences, i.e. repeats as genome structure targets (IR1, IR2, IR4), proteins involved in transformation (LMP1, EBNA3C) and in maintenance of latency (EBNA1). Upon genome destruction, they further noted proliferation arrest and apoptosis of EBV^+ cells, but not cytotoxicity. Yuen et al., 2015 performed targeted editing of the EBV genome in different EBV-infected human cell lines using two gRNAs to create a virus not expressing BART miRNA by excising a fragment of the BART promoter. The authors then expanded this study to other EBV targets (EBNA1, IR1, oriP) in the NPC cell line C666-1 and could show a decrease in EBV load, a progressive but incomplete EBV supression and that cells were sensitised to chemotherapy (Yuen et al., 2017). A third team showed complete inhibition of viral replication and in some cases elimination of the genome from infected cells in three different herpesviruses (HSV-1, CMV, EBV) (van Diemen et al., 2016).

These studies have shown the applicability of this new technology of genome editing to EBV. However, for therapeutic purposes, approaches for delivery need to be developed. Adenovirus as used in gene therapy has been suggested as a promising method. Furthermore, engineering delivery viruses that mimic the cell-tropism of the target virus could be an option (Wang and Quake, 2014).

1.7 EBV genome

1.7.1 Genome structure

The genome of EBV is a circular, double-standed DNA molecule of around 172 kb length (figure 1.3). The GC content varies accross the genome, but is on average 57%. Compared to Herpes simplex virus 1 (HSV-1), the GC content is lower. This is thought to be a result of EBV establishing latency in a dividing cell population compared to non-dividing neurons in the case of HSV-1. As EBV has to replicate when the host cell divides, 5-methylcytosine residues are spontaneously deaminated to thymidine and fixed during replication (Honess et al., 1989).

Repeat regions

The genome of EBV (and other herpesviruses) contains several repeat regions. In terms of herpesvirus biology, the genome can be classified structurally as a class C genome: It has a terminal repeat region of 500 bp segments. Additionally, there are four internal direct repeats and a number of smaller repeats (for example the family of repeats (FR) and direct repeats (DR)). The internal repeat 1 (IR1) divides the genome into two unique regions (unique short, U_S and unique long, U_L), while the other three (IR2, IR3 and IR4) lie within the U_L region. In contrast to other Gammaherpesvirinae such as cottontail

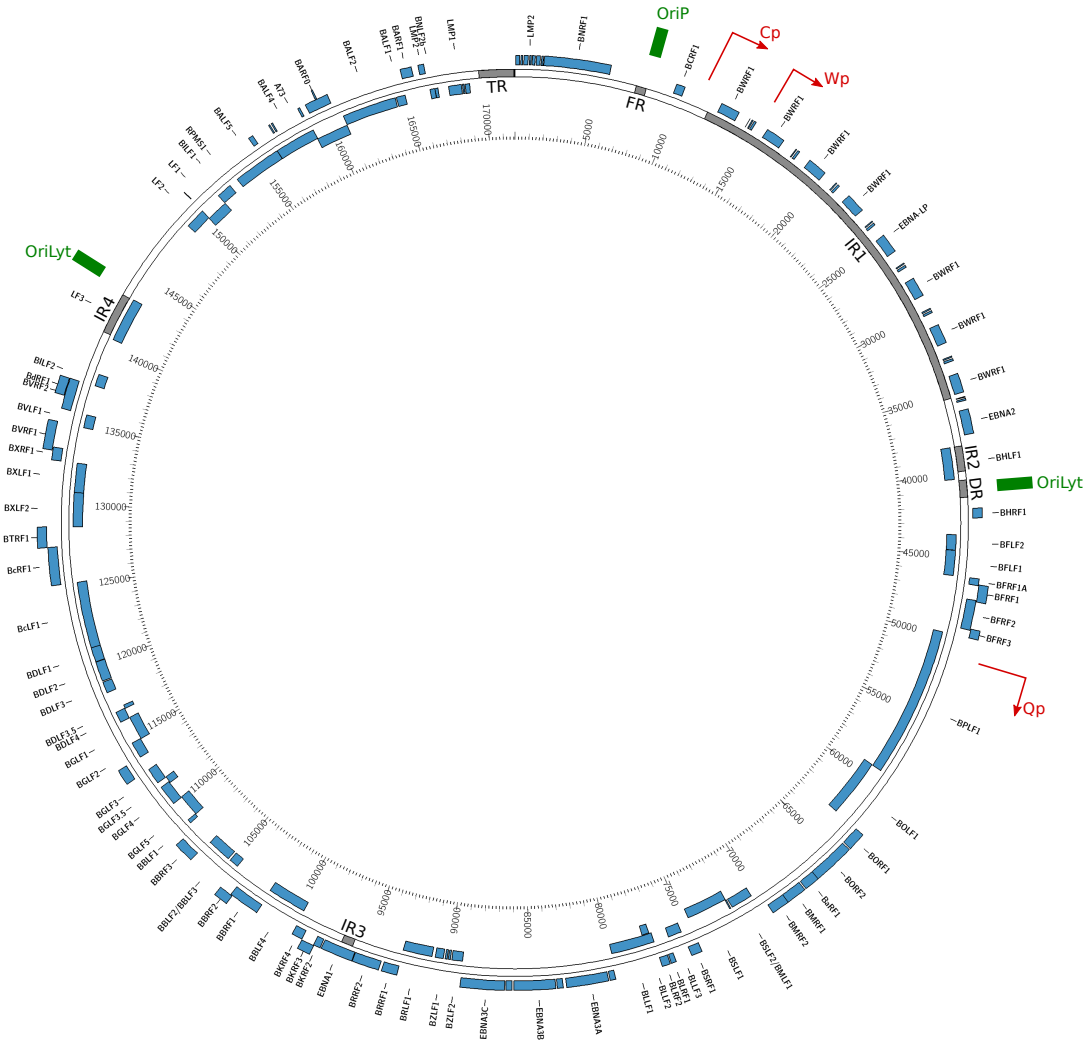


FIGURE 1.3: Circular map of the genome. Blue: Open reading frames (outer track: forward, inner track: reverse); Grey: Repeat regions; Red: promoters; Green: origins of replication; Pink: non-coding RNAs.

rabbit herpesvirus, segment inversion cannot occur as the internal and terminal repeats are unrelated (Davison, 2007; Lacoste et al., 2010). The copy number of internal and terminal repeats can vary between isolates, as a result genome length is variable.

The role of most of the repeats is probably related to replication (Davison, 2007). The terminal repeats are part of the mechanism to circularise the EBV genome after infection of the host cell and are essential for the DNA packaging during virion production (Zimmermann and Hammerschmidt, 1995).

IR1 has a particular functional role, as it contains the viral Wp promoter which is used to transcribe the *EBNA* genes. It is made up of five to ten copies of a 3072 bp long W segment (repeat unit). Additionally, every repeat unit contains two exons of *EBNA-LP*. It is therefore important in EBV's transforming abilities as its length determines the number

of available Wp promoters as well as the size of the transforming latency protein EBNA-LP. Most indviduals carry virus with five to eight repeat units and it has been shown that at least five copies are required for optimal B cell transformation, while fewer copies lead to a progressively decreasing ability to transform (Tierney et al., 2011).

Smaller repeats may also play in role in regulation of translation, for example via self-inhibition of synthesis. Such a repeat, IR3, consisting of Gly-Ala residues (GAr) of variable length lies between two Gly-Arg rich regions at the N-terminus of EBNA1. It has been shown to alter MHC-I presentation of EBNA1-derived peptides to $CD8^+$ T cells by suppressing mRNA translation in cis (Yin, Manoury, and Fåhraeus, 2003). This mechanism is likely due to the purine-rich mRNA sequence, rather than the encoded GAr polypeptide (Tellam et al., 2012).

The origin of replication used during latency, OriP, contains the family of repeats (FR) region and the dyad symmetry (DS) element, which are binding sites for EBNA1 (Yates, Camiolo, and Bashaw, 2000). FR is the plasmid maintenance element and is tethered via bound EBNA1 to condensed mitotic chromosomes, thereby ensuring the segregation of plasmids to both daughter cells (Hung, Kang, and Kieff, 2001). DS is the origin of replication and is the site at which EBNA1-dependent bidirectional DNA synthesis starts (Hammerschmidt and Sugden, 2013).

Open reading frames and promoters

Spliced genes are rather uncommon in EBV, and apart from a few exceptions (the EBNA genes are spliced from one 5'-leader), every gene has its own promoter. Moreover, ORFs are not particularly clustered according to function or expression kinetics (Davison, 2007).

Latent proteins are expressed under the control of three promoters: Wp, Qp, and Cp (see figure 1.3). Wp and Cp are located in the BamHI W fragment, with Wp being the first promoter to be activated during *in vitro* immortalisation. Very early during infection, a switch from Wp to Cp occurs due to transactivation from EBNA1 and -2 on Cp (Woisetschlaeger et al., 1990). Another promoter, Qp, located in the BamHI Q fragment, drives EBNA1 expression and has been shown to be constitutively active in various EBV^+ cell lines and tumours (Tao et al., 1998). It was found to be consistently hypomethylated, resulting in the expression of EBNA1 no matter the cell type, type of latency, or the activity of other promoters (Tao et al., 1998).

Non-coding RNAs

EBV-encoded small RNAs The most abundant non-coding (nc)RNAs in EBV are the EBV-encoded small RNAs (EBERs) (Rymo, 1979). They are therefore an excellent target to identify EBV-infected cells in tissue using *in situ* hybridisation (Ambinder and Mann, 1994). There are two transcripts, EBER1 and EBER2, which are 167 bp and 172 bp long, respectively, and which are divided by a 161 bp spacer region. Even though they only share 54 % sequence similarity, both EBERs form stable stem-loop structures very similar to each other (Rosa et al., 1981). EBER expression seems to be restricted to latently

infected cells, whereas tissue associated with lytic replication of EBV, such as oral hairy leukoplakia, is EBER-free (Gilligan et al., 1990).

EBERs are not necessary for B cell transformation (Swaminathan, Tomkinson, and Kieff, 1991), but contribute to it (Yajima, Kanda, and Takada, 2005). They are thought to play a role in cancerogenesis, as they are necessary to maintain the malignant phenotype of BL cells (Komano et al., 1999) and provide apoptotic resistance in BL (Nanbo et al., 2002) as well as epithelial cells (Nanbo, Yoshiyama, and Takada, 2005). They induce a number of cytokines that act as autocrine gowth factors, including IL-10 in BL cells, IGF-1 in NPC and EBVaGC cells, and IL-9 in T cells (Iwakiri and Takada, 2010). Additionally, they have been shown to interact with and activate the innate immune system (Iwakiri et al., 2009).

BamHI A rightward transcripts The BamHI A rightward region also encodes a number of transcripts (BARTs). The BART region contains seven exons named I to VII, including some alternative variants, which are alternatively spliced. There are at least six BART splicing forms – *BARF0*, *RK-BARF0*, *RPMS1*, *RPMS1A A73*, and *RB3* – each of them containing a putative ORF. Expression of native proteins, however, have rarely been shown conclusively *in vivo* (Yamamoto and Iwatsuki, 2012). Their specific function remains unknown.

BamHI A rightward transcript miRNAs The intronic regions of the BARTs encode clusters of microRNAs (BART-miRNAs), small non-coding RNAs ranging in size from 20 to 25 bp. There are two BART-miRNA clusters, comprising of 44 verfied mature miRNAs derived from 22 precursors (Young, Yap, and Murray, 2016). Viruses lacking this locus are only moderately impaired in their transforming ability (Vereide et al., 2014; Kanda et al., 2015). In fact, it is to a large extent deleted in the strain B95-8. However, BART-miRNAs have been suggested to play a role in oncogenesis, as some of them provide an anti-apoptotic effect (Kang and Kieff, 2015; Vereide et al., 2014) and have been shown to promote tumour growth *in vivo* (Qiu et al., 2015). They further maintain latency by targeting lytic genes (Pfeffer et al., 2004; Barth et al., 2008) and modulate LMP1 expression for immune evasion (Lo et al., 2007). Interestingly, miRNA might not only target the EBV-infected cell in which they are produced. The release of exosomes containing miRNAs as well as LMP1 has been observed in NPC cells (Meckes et al., 2010; Gourzones et al., 2010), and the miRNAs were able to affect neighbouring cells. As the exosomes were found in the serum samples, they can likely even reach more distant cells (Gourzones et al., 2010).

BamHI H rightward transcript miRNAs The BamHI H rightward fragment is the second region in the EBV genome encoding a cluster of three miRNAs (BHRF1-miRNAs) (Pfeffer et al., 2004), which are highly expressed in latency III expressing LCLs (Cai et al., 2006; Cosmopoulos et al., 2009). Viruses lacking the BHRF1-miRNA locus are 20 times less efficient in B cell transformation than wild-type (Feederle et al., 2011a) showing that it contributes largely to the transforming capacity of EBV (Feederle et al., 2011a; Feederle

et al., 2011b; Vereide et al., 2014). However, in tumour samples of NPC, GC and DLBCL patients, BHRF1-miRNAs are largely absent (Chen et al., 2010b; Marquitz et al., 2013; Imig et al., 2011). In contrast to that, AIDS-related DLBCL and BL showed high BHRF1-miRNA expression (Xia et al., 2008). They also contribute to immune evasion, e.g. by targeting CXCL11, a T cell attracting chemokine (Xia et al., 2008).

1.7.2 Genetic diversity in EBV

There is a great interest in identifying variants that are associated with specific diseases, especially since incidence rates of associated tumours vary across the world.

Types of EBV

Based on the sequence variation found in the EBNA2 and EBNA3A-C genes, there two types defined (Sample et al., 1990). On the amino acid level, the two types differ at around 53 % of positions in EBNA2 and 16-28 % in EBNA3A, -B, and -C (Sample et al., 1990).

Type 1 and 2 are equally prevalent in Africa, but in the rest of the world, type 1 is the predominant strain (Zimber et al., 1986; Young et al., 1987). Type 2 is less effective in infecting B cells, and type 1-infected LCLs show a superior growth (Rickinson, Young, and Rowe, 1987). A single amino acid substitution in EBNA2 has been found to be responsible for this difference (Tzellos et al., 2014).

In a number of patient cohorts, in particular HIV-positive and other immunosuppressed patients, type 1 and 2 superinfections have been demonstrated (Sculley et al., 1990; Srivastava et al., 2000; Santón et al., 2011). This allows the possibility of recombination occuring between types, and indeed, naturally occurring intertypic recombinants have been described (Midgley et al., 2000; Yao et al., 1996; Kim, Kang, and Lee, 2006; Palser et al., 2015).

Gene polymorphisms

LMP1 LMP1 is an oncogene which funtionally mimics a constitutively active form of CD40 and is responsible for phenotypic changes in both B cells and epithelial cells. The protein is encoded within three exons. It consists of an N-terminal domain of 24 aa, six transmembrane domains each 20 aa long, and a C-terminal domain from position 185 to 386, which contains two C-terminal activation regions (CTAR1 and -2) (*Uniprot Database*).

Given the functional importance of the C-terminus, many studies looking at LMP1 diversity have focused on this region. Comparisons of polymorphisms in the C-terminus has led to the classification of eight LMP1 variants: the reference strain B95-8, China 1, China 2, China 3, North Carolina (NC), Mediterranean+, Mediterranean-, and Alaskan (Ai et al., 2012; Edwards, Seillier-Moiseiwitsch, and Raab-Traub, 1999; Mainou and Raab-Traub, 2006).

A commonly reported variation observed in LMP1 is a 30 bp deletion in the C-terminus (del-LMP1), which results in the loss of amino acids 343-352 (Miller et al., 1994). This variation can be found worldwide (Correa et al., 2004; da Costa, Marques-Silva, and Moreli,

2015; Lorenzetti et al., 2012) and it has been shown to occur in various EBV-associated malignancies, such as NPC (da Costa, Marques-Silva, and Moreli, 2015; Banko et al., 2016), HL (Ai et al., 2012; Guiretti et al., 2007; Zhou et al., 2001; Lorenzetti et al., 2012), EBVaGC (Chen et al., 2010a; Chen et al., 2011; BenAyed-Guerfali et al., 2011), as well as IM (Ai et al., 2012; Lorenzetti et al., 2012; Berger et al., 1997). The variant is therefore thought to increase oncogenicity (Ai et al., 2012; Tao et al., 1998; Zhang et al., 2002).

The ten deleted amino acids affect the transformation effector site 2 (TES2) as well as the CTAR2 region which plays a role in NF-κB activation. The transformation potential seems to be enhanced in del-LMP1 carrying viruses compared to restored wild type (WT)-LMP1 (Li, Chang, and Liu, 1996). However, not all studies could confirm a functional difference for this variant (Fielding et al., 2001; Yeh et al., 1997) and the issue remains controversial.

Another frequent variant is the substitution of G169425T at the XhoI restriction site at the N-terminus of *LMP1*, which results in its loss (Hu et al., 1991). The biological significance and consequence is uncertain, but due to its high prevelance in the Asian population where NPC is endemic, the association was made between XhoI-loss and the development of NPC (da Costa, Marques-Silva, and Moreli, 2015). However, in NPC cases in North Africa, this variant is not observed (Ayadi et al., 2007; Bouzid et al., 1994). Similarly, EBVaGC cases in China often display the XhoI-loss (Chen et al., 2010a), while studies from other geographic regions do not (Abdirad et al., 2007; Ordonez et al., 2011). One study of Asian populations has also associated XhoI-loss with a number of other malignancies such as IM, hemophagocytic lymphohistiocytosis, and HL (Ai et al., 2012). As most studies are biased towards sampling Asian populations, some papers have suggested this variant is likely only a geographic variation (Ai et al., 2012).

LMP2A Variation in LMP2A has been described, but polymorphisms have not been associated with any specific diseases (Tanaka et al., 1999; Berger et al., 1999; Wang et al., 2010b; Han et al., 2012).

EBNA1 EBNA1 variants have mostly been classified based on their amino acid polymorphism at the C-terminus. The prototype sequence is P-ala (which includes the WT strain) and variant strains are V-val, V-leu, V-pro, P-thr (Bhatia et al., 1996).

Many authors have found that the V-val variant is most commonly found in Asia (Gutiérrez et al., 1997; Mai et al., 2007; Zhang et al., 2004; Sandvej, Zhou, and Hamilton-Dutoit, 2000). It was suggested that the V-val variant is associated with development of NPC (Zhang et al., 2004; Mai et al., 2007), due to expression differences of viral and host genes as well as itself in epithelial cells (Mai et al., 2007). V-val was also reported to have a higher transcriptional activity than P-ala (WT), and a higher binding affinity for the FR element, which among other roles is a transcription enhancer for the LMP promoter as well as the Cp promoter (Mai et al., 2010; Zhang et al., 2004). Another study showed that the V-val variant might have a stronger anti-apoptotic effect on cells deprived of serum and therefore growth-signals, a hallmark for tumour cells (Chao et al., 2015).

However, other studies have weakened the suggested strong association with NPC, as it was found to be most common in both healthy carriers as well as NPC, EBVaGC, IM, HL and Hemophagocytic lymphohistiocytosis patients in Asia (Do et al., 2008; Chang et al., 2009; Ai et al., 2012; Wang et al., 2010d). A recent study analysed EBV gene variation in NPC biopsies in Caucasians from Serbia, a non-endemic region. They found EBNA1 P-thr and P-ala to be the most prevalent types, confirming the geographic association of V-val with Asia (Banko et al., 2016). Further, they found a particular subtype of P-thr to be associated with NPC type III in their cohort and proposed a particular combination of EBNA2, LMP1 and EBNA1 polymorphisms (type1/Med/P-thr) to be a risk factor for developing advanced NPC stages.

EBNA1 plays an important role in oncogenesis but evidence for disease association is not clear. Furthermore, there is a clear bias in the literature of studying populations from Asia. As a consequence, disentangling the effect of geography and disease association is difficult. It requires careful studies with large sample sizes in well defined subpopulations and controls.

EBNA2 The type-specific variations found in EBNA2 have not been found to be associated with a specific disease (Tzellos and Farrell, 2012). However, type 2 has been found to be less efficient in transformation due to a single amino acid change in the transactivation domain, S442D (Tzellos et al., 2014).

EBNA3A, -3B, -3C Variations differentiating between type 1 and 2 in the *EBNA2* gene are linked to the polymorphisms in the *EBNA3* genes (Rowe et al., 1989). Six EBNA3 variants, including the prototype WT (B95.8), were defined based on linked polymorphisms in the *EBNA3A*, *-B* and *-C* genes. Recombination of variation patterns was also observed and formed the most distinct type (Görzer et al., 2006). Evidence for disease association of polymorphisms, however, is weak. In a study of *EBNA3C* variation in northern China, no association between variation and EBVaGC, NPC, or asymptomatic carriers could be found (Wu et al., 2012). However, it has been shown that EBNA3B, which acts as a tumour suppressor, carrying loss of function mutants leads to increased transforming ability and promotes formation of DLBCL-like tumours in humanised mice, partly due to reduced T-cell mediated killing *in vivo* (White et al., 2012). Similar phenotypic tumours in humans were also found to carry a truncated EBNA3B, and cell lines derived from these lymphoma displayed a similar gene expression profile as the tumour-derived lines from the humanised mice.

BZLF1 *BZLF1* encodes the major switch from latent to lytic cycle Zta or ZEBRA. It is therefore of particular interest whether and how variation in the protein itself and within the promoter impacts function and/or expression. There are now four variant types of the gene defined based on differences to the prototype WT (B95.8): A, B, C (Luo et al., 2011) and D (Lorenzetti et al., 2014). Whether or not *BZLF1* variation corresponds to

pathogenesis or whether it is geographically restricted is still controversial. In NPC patients from Tunisia, one study found BZLF1 variants differentially distributed between tumour cells and lymphocytes (Sacaze et al., 2001). In a study from 2011, no association between *BZLF1* variations and EBVaGC or NPC in China was found (Luo et al., 2011), but a later study showed that BZLF1 type B occurs slightly more often in lymphoma compared to EBVaGC or NPC (Yang et al., 2014). In contrast to that, Lorenzetti et al., 2014 found a subtype of BZLF1-A particularly frequent in lymphoma and BZLF1-C in IM in patients from Argentina.

The promoter has also been classified into four types: Zp-P, which is identical to the WT, as well as Zp-V1, Zp-V3 and Zp-V4 based on four positions (-196, -141, -106 and -100) (Lorenzetti et al., 2009). First studies found a differential distribution of promoter variants between malignant and benign EBV infections (Gutiérrez et al., 2002), but later, malignancy associated variants such as Zp-V3 were also found in healthy carriers and IM patients (Tong et al., 2003; Martini et al., 2007; Imajoh et al., 2012). In particular Zp-V3 was also associated with severe diseases such as chronic active EBV (CAEBV) (Jin et al., 2010; Imajoh et al., 2012). However, functional studies to confirm potential associations are lacking.

Virus strains

In EBV, most genes have only a low level of natural diversity on the amino acid level (less than 5 %). The most variable genes are the latency genes (within and between EBV type variation) (Chang et al., 2009; Palser et al., 2015) and many studies have focused on these and other polymorphic genes to study variation. But this approach gives an incomplete picture of variation, as information of linkage between variable sites and potential larger scale rearrangements of genomic regions or recombination cannot be observed.

EBV genome sequencing The first complete EBV genome, strain B95-8, was published in 1984 (Baer et al., 1984). The virus originated from the 883L cell line, which was established by culturing B cells from a North American IM patient. This virus was used to infect a marmoset B cell line, from which DNA was subsequently isolated and sequenced by Sanger sequencing of cloned EcoRI and BamHI fragments (Baer et al., 1984). The genome has since been updated by using it as backbone and filling an atypical large deletion in the BamHI A fragment with a sequence from the strain Raji (de Jesus, 2003). Only for this historical reason is this sequence often referred to as the wild type (WT) or reference sequence.

The first – and up until recently only – type 2 genome, AG876, was published almost 20 years later and originated from a BL case in Ghana (Dolan et al., 2006). The sequence was obtained by digesting the genome into fragments which were then cloned and Sanger sequenced. Genome comparison of the WT sequence and AG876 confirmed how similar the genomic sequence of both types is apart from the typing genes.

In 2005, the Chinese strain GD1 from the southern Guangdong province was sequenced. The virus was isolated from the saliva of a NPC patient and used to infect

umbilical cord mononuclear cells (Zeng et al., 2005). Here, the isolated DNA was then PCR-amplified, cloned and Sanger sequenced. Six years later, another Guangdong strain, GD2, was published. This genome is the first one derived from virus directly isolated from clinical material (a NPC tumour) and sequenced using Next generation sequencing (NGS) (Liu et al., 2011).

Due to the advancements of NGS technology, GD2 also marks the beginning of the increase in EBV genome sequencing: HKNPC1, a NPC sample from Hong Kong in 2012 (Kwok et al., 2012); Akata, a Japanese BL cell line, and Mutu, a Kenyan BL cell line in 2012 (Lin et al., 2012); K4123-Mi and K4413-Mi, spontaneous LCLs (sLCLs) generated from healthy North American donors, in 2013 (Lei et al., 2013); C666-1, a Chinese NPC cell line, in 2013 (Tso et al., 2013); and M81, a Chinese NPC isolate, in 2013 (Tsai et al., 2013). Moreover, due to EBV's prevalence in the human population, one study used sequencing raw data from the 1000 Human Genomes Project to filter out EBV-specific reads and assemble full genomes (Santpere et al., 2014).

A further advancement in EBV sequencing occurred in 2014, when eight further NPC samples from Hong Kong (HKNPC2 to -9) were published (Kwok et al., 2014). The previous sequencing attempts were tempered by the fact that DNA isolated from clinical material and even cell lines only contained a very small percentage of virally derived material. Kwok et al., 2014 were the first to publish EBV genomes obtained using a recently developed targeted enrichment technique for pathogen genomics (Depledge et al., 2011). In this method, isolated DNA is sheared and then hybridised with biotinylated RNA baits which are complementary to sequences of interest. The baits are then pulled out, bound to the sample sequences of interest, and after digesting the RNA, the enriched sample can be sequenced. Using this method, Palser et al., 2015 then sequenced a big number of 71 new genomes from various tumour types and normal infections of different geographic origins, including eleven additional type 2 sequences. Since then more publications came out using this method (Liu et al., 2016; Chiara et al., 2016)

Table 1.2 lists all available whole genome sequences to date (April 2017).

There are phenotypic difference between strains: for example, the deletion in B95-8 affects one of the lytic origins of replication (Baer et al., 1984). Nevertheless, it is able to establish LCLs from peripheral B cells. M81, an NPC isolate from Hong Kong, has been found to have a reversed tropism relative to other known strains, i.e. it predominantly infects epithelial cells probably due to an abundance of gp110 (Tsai et al., 2013).

Considering the whole genome, the most polymorphic genes are the latency genes. A particularly high number of nonsynonymous changes are observed in *EBNA2* and the *EBNA3* family, as well as *LMP1*. Among the lytic genes, the most diversity is found in *BDLF3* and *BLLF1* (encoding for gp150 and gp350), *BZLF1* (Zta), *BRRF2* (tegument protein), and *BNLF2a* (plays a role in immune evasion) (Palser et al., 2015). These genes are likely partly or as a whole under positive selection due to being immunogenic.

Comparative analysis of whole genomes has also elucidated that recombination is

a frequent event in EBV. Palser et al., 2015 described both inter- and intratypic recombinants, as did other publications, in particular in geographically constrained genomes (Kwok et al., 2014; Santpere et al., 2014).

Name	Type	Geographic origin	EBNA2 type	Accession no.	Reference	
WT	IM	USA	1	NC_007605	(Baer et al., 1984, de Jesus, 2003)	●
AG876	BL	Ghana	2	NC_009334	(Dolan et al., 2006)	
GD1	NPC—saliva	China	1	AY961628	(Zeng et al., 2005)	●
GD2	NPC—tumor	China	1	HQ020558	(Liu et al., 2011)	●
HKNPC1	NPC—tumor	Hong Kong	1	JQ009376	(Kwok et al., 2012)	●
Akata	BL	Japan	1	KC207813	(Lin et al., 2012)	●
Mutu	BL	Kenya	1	KC207814	(Lin et al., 2012)	●
K4123-Mi	sLCL	USA	1	KC440851	(Lei et al., 2013)	●
K4413-Mi	sLCL	USA	1	KC440852	(Lei et al., 2013)	●
M81	NPC	Hong Kong	1	KF373730	(Tsai et al., 2013)	●
HKNPC2	NPC	Hong Kong	1	KF992564	(Kwok et al., 2014)	●
HKNPC3	NPC biopsy	Hong Kong	1	KF992565	(Kwok et al., 2014)	●
HKNPC4	NPC biopsy	Hong Kong	1	KF992566	(Kwok et al., 2014)	●
HKNPC5	NPC biopsy	Hong Kong	1	KF992567	(Kwok et al., 2014)	●
HKNPC6	NPC biopsy	Hong Kong	1	KF992568	(Kwok et al., 2014)	●
HKNPC7	NPC biopsy	Hong Kong	1	KF992569	(Kwok et al., 2014)	●
HKNPC8	NPC biopsy	Hong Kong	1	KF992570	(Kwok et al., 2014)	●
HKNPC9	NPC biopsy	Hong Kong	1	KF992571	(Kwok et al., 2014)	●
Saliva1	Healthy saliva	UK	1	LN824142	(Palser et al., 2015)	●
HL01	HL	UK	1	LN824226	(Palser et al., 2015)	●
HL02	HL	UK	1	LN827546	(Palser et al., 2015)	●
HL04	HL	UK	1	LN827564	(Palser et al., 2015)	●
HL05	HL	UK	1	LN824204	(Palser et al., 2015)	●
HL08	HL	UK	1	LN824225	(Palser et al., 2015)	●
HL09	HL	UK	1	LN827522	(Palser et al., 2015)	●
HL11	HL	UK	1	LN827524	(Palser et al., 2015)	●
L591	HL cell line	Germany	1	LN827523	(Palser et al., 2015)	●
YCCEL1	GC cell line	South Korea	1	LN827561	(Palser et al., 2015)	●
HKN14	sLCL	Hong Kong	1	LN824209	(Palser et al., 2015)	●
HKN15	sLCL	Hong Kong	1	LN827547	(Palser et al., 2015)	●
HKN19	sLCL	Hong Kong	1	LN824224	(Palser et al., 2015)	●
D3201.2	NPC biopsy	China	1	LN827549	(Palser et al., 2015)	●
C666-1 resequence	NPC cell line	China	1	LN827525	(Palser et al., 2015)	●
M-ABA	LCL, NPC virus	N. Africa	1	LN827527	(Palser et al., 2015)	●
Daudi	BL	Kenya	1	LN827545	(Palser et al., 2015)	●
BL36	BL	N. Africa	2	LN827557	(Palser et al., 2015)	
BL37	BL	Africa	1	LN827526	(Palser et al., 2015)	●
Makau	BL	Kenya	1	LN827551	(Palser et al., 2015)	●
Mak1 duplicate	BL	Kenya	1	LN824203	(Palser et al., 2015)	●
sLCL-IM1.02	sLCL, IM	Australia	1	LN827596	(Palser et al., 2015)	●
sLCL-IM1.05	sLCL, IM	Australia	1	LN827590	(Palser et al., 2015)	●
sLCL-IM1.09	sLCL, IM	Australia	1	LN827567	(Palser et al., 2015)	●
sLCL-IM1.16	sLCL, IM	Australia	1	LN827799	(Palser et al., 2015)	●
sLCL-IM1.17	sLCL, IM	Australia	1	LN827583	(Palser et al., 2015)	●
sLCL-IS1.01	sLCL, PTLD	Australia	1	LN827570	(Palser et al., 2015)	●
sLCL-IS1.03	sLCL, PTLD	Australia	1	LN827595	(Palser et al., 2015)	●
sLCL-IS1.04	sLCL, PTLD	Australia	1	LN827597	(Palser et al., 2015)	●
sLCL-IS1.06	sLCL, PTLD	Australia	1	LN827584	(Palser et al., 2015)	●
sLCL-IS1.07	sLCL, PTLD	Australia	1	LN827594	(Palser et al., 2015)	●
sLCL-IS1.08	sLCL, PTLD	Australia	1	LN827553	(Palser et al., 2015)	●
sLCL-IS1.10	sLCL, PTLD	Australia	1	LN827592	(Palser et al., 2015)	●

Name	Type	Geographic origin	EBNA2 type	Accession no.	Reference	
sLCL-IS1.11	sLCL, PTLD	Australia	1	LN827569	(Palser et al., 2015)	●
sLCL-IS1.12	sLCL, PTLD	Australia	1	LN827593	(Palser et al., 2015)	●
sLCL-IS1.13	sLCL, PTLD	Australia	1	LN827578	(Palser et al., 2015)	●
sLCL-IS1.14	sLCL, PTLD	Australia	1	LN827575	(Palser et al., 2015)	●
sLCL-IS1.15	sLCL, PTLD	Australia	1	LN827586	(Palser et al., 2015)	●
sLCL-IS1.18	sLCL, PTLD	Australia	1	LN827572	(Palser et al., 2015)	●
sLCL-IS1.19	sLCL, PTLD	Australia	1	LN827588	(Palser et al., 2015)	●
sLCL-IS1.20	sLCL, PTLD	Australia	1	LN827576	(Palser et al., 2015)	●
sLCL-1.02	sLCL	Kenya	1	LN827558	(Palser et al., 2015)	●
sLCL-BL1.03	sLCL	Kenya	1	LN827582	(Palser et al., 2015)	●
sLCL-1.04	sLCL	Kenya	1	LN827585	(Palser et al., 2015)	●
sLCL-1.05	sLCL	Kenya	1	LN827581	(Palser et al., 2015)	●
sLCL-1.06	sLCL	Kenya	1	LN827566	(Palser et al., 2015)	●
sLCL-1.07	sLCL	Kenya	1	LN827565	(Palser et al., 2015)	●
sLCL-1.08	sLCL	Kenya	1	LN827552	(Palser et al., 2015)	●
sLCL-1.09	sLCL	Kenya	1	LN827574	(Palser et al., 2015)	●
sLCL-1.10	sLCL	Kenya	1	LN827573	(Palser et al., 2015)	●
sLCL-1.11	sLCL	Kenya	1	LN827550	(Palser et al., 2015)	●
sLCL-1.12	sLCL	Kenya	1	LN824205	(Palser et al., 2015)	●
sLCL-1.13	sLCL	Kenya	1	LN827579	(Palser et al., 2015)	●
sLCL-1.17	sLCL	Kenya	1	LN827577	(Palser et al., 2015)	●
sLCL-1.19	sLCL	Kenya	1	LN827562	(Palser et al., 2015)	●
sLCL-BL1.20	sLCL	Kenya	1	LN827571	(Palser et al., 2015)	●
sLCL-1.24	sLCL	Kenya	1	LN827568	(Palser et al., 2015)	●
pLCL-TRL1-pre[1]	sLCL, PTLD	USA	1	LN824207	(Palser et al., 2015)	●
pLCL-TRL1-post[1]	sLCL, PTLD	USA	1	LN8242076	(Palser et al., 2015)	
pLCL-TRL595	sLCL, PTLD	USA	1	LN827559	(Palser et al., 2015)	●
X50-7	LCL	USA	1	LN827555	(Palser et al., 2015)	●
Wewak1	BL	PNG	2	LN827544	(Palser et al., 2015)	
AFB1	LCL	Unknown	2	LN827554	(Palser et al., 2015)	
Jijoye	BL	Nigeria	2	LN827800	(Palser et al., 2015)	
P3HR1 c16	BL	Nigeria	2	LN827548	(Palser et al., 2015)	
Cheptages	BL	Kenya	2	LN827556	(Palser et al., 2015)	
sLCL-IS2.01	sLCL, PTLD	Australia	2	LN827589	(Palser et al., 2015)	
sLCL-2.14	sLCL	Kenya	2	LN827560	(Palser et al., 2015)	
sLCL-2.15	sLCL	Kenya	2	LN827591	(Palser et al., 2015)	
sLCL-2.16	sLCL	Kenya	2	LN827580	(Palser et al., 2015)	
sLCL-1.18	sLCL	Kenya	2	LN827563	(Palser et al., 2015)	
sLCL-2.21	sLCL	Kenya	2	LN827587	(Palser et al., 2015)	
sLCL-2.22	sLCL	Kenya	2	LN831023	(Palser et al., 2015)	
NA12878	Healthy*	Europe	1	NA	(Lei et al., 2013)	
NA19114	Healthy*	Africa	1	NA	(Santpere et al., 2014)	
NA19315	Healthy*	Africa	1	NA	(Santpere et al., 2014)	
NA19384	Healthy*	Africa	1	NA	(Santpere et al., 2014)	
EBVaGC1	EBVaGC biopsy	China	1	KT273942	(Liu et al., 2016)	
EBVaGC2	EBVaGC biopsy	China	1	KT273943	(Liu et al., 2016)	
EBVaGC3	EBVaGC biopsy	China	1	KT254013	(Liu et al., 2016)	
EBVaGC4	EBVaGC biopsy	China	1	KT273944	(Liu et al., 2016)	
EBVaGC5	EBVaGC biopsy	China	1	KT273945	(Liu et al., 2016)	
EBVaGC6	EBVaGC biopsy	China	1	KT273946	(Liu et al., 2016)	
EBVaGC7	EBVaGC biopsy	China	1	KT273947	(Liu et al., 2016)	
EBVaGC8	EBVaGC biopsy	China	1	KT273948	(Liu et al., 2016)	
EBVaGC9	EBVaGC biopsy	China	1	KT273949	(Liu et al., 2016)	
LC1	LC biopsy	China	1	KT823506	(Wang et al., 2016)	
LC2	LC biopsy	China	1	KT823507	(Wang et al., 2016)	
LC3	LC biopsy	China	1	KT823508	(Wang et al., 2016)	
LC4	LC biopsy	China	1	KT823509	(Wang et al., 2016)	

Name	Type	Geographic origin	EBNA2 type	Accession no.	Reference
MP	BL biopsy	Brazil	1	KP968258	(Lei et al., 2015)
SCL	BL biopsy	Brazil	1	KP968259	(Lei et al., 2015)
CCH	BL biopsy	Brazil	1	KP968257	(Lei et al., 2015)
H018436D	BL biopsy	Ghana	1	KP968262	(Lei et al., 2015)
H058015C	BL biopsy	Ghana	1	KP968263	(Lei et al., 2015)
HU11393[1]	BL biopsy	Ghana	1	KP968261	(Lei et al., 2015)
H03753A[1]	BL biopsy	Ghana	1	KR063342	(Lei et al., 2015)
H002213	BL biopsy	Ghana	1	KP968264	(Lei et al., 2015)
CV- ARG	BL biopsy	Argentina	1	KR06 3343	(Lei et al., 2015)
RPF	BL biopsy	Brazil	1	KR06 3344	(Lei et al., 2015)
FNR	BL biopsy	Brazil	1	KR06 3345	(Lei et al., 2015)
VGO	BL biopsy	Brazil	1	KP968260	(Lei et al., 2015)
CAR	sLCL, Healthy	Italy	1	ERS1100719	(Chiara et al., 2016)
PP	sLCL, MS	Italy	1	ERS1100731	(Chiara et al., 2016)
MV	sLCL, MS	Italy	1	ERS1100733	(Chiara et al., 2016)
BL	sLCL, MS	Italy	1	ERS1100735	(Chiara et al., 2016)
VL	sLCL, MS	Italy	1	ERS1100730	(Chiara et al., 2016)
NM	sLCL, Healthy	Italy	1	ERS1100715	(Chiara et al., 2016)
MC	sLCL, Healthy	Italy	1	ERS1100718	(Chiara et al., 2016)
MFA	sLCL, MS	Italy	1	ERS1100723	(Chiara et al., 2016)
GV	sLCL, MS	Italy	1	ERS1100726	(Chiara et al., 2016)
GIOVS	sLCL, Healthy	Italy	1	ERS1100717	(Chiara et al., 2016)
CS	sLCL, MS	Italy	1	ERS1100724	(Chiara et al., 2016)
GR	sLCL, Healthy	Italy	1	ERS1100714	(Chiara et al., 2016)
PT	sLCL, Healthy	Italy	1	ERS1100716	(Chiara et al., 2016)
BA	sLCL, MS	Italy	1	ERS1100728	(Chiara et al., 2016)
MM	sLCL, MS	Italy	1	ERS1100734	(Chiara et al., 2016)
LOL	sLCL, MS	Italy	1	ERS1100725	(Chiara et al., 2016)
IM	sLCL, MS	Italy	1	ERS1100727	(Chiara et al., 2016)
GF	sLCL, MS	Italy	1	ERS1100729	(Chiara et al., 2016)
BR	sLCL, Healthy	Italy	1	ERS1100721	(Chiara et al., 2016)
CM	sLCL, MS	Italy	1	ERS1100732	(Chiara et al., 2016)
LUL	sLCL, Healthy	Italy	1	ERS1100722	(Chiara et al., 2016)
TM	sLCL, MS	Italy	1	ERS1100713	(Chiara et al., 2016)
SC	sLCL, MS	Italy	1	ERS1100710	(Chiara et al., 2016)
SA	sLCL, MS	Italy	1	ERS1100711	(Chiara et al., 2016)
RT	sLCL, MS	Italy	1	ERS1100712	(Chiara et al., 2016)
MST	sLCL, Healthy	Italy	1	ERS1100720	(Chiara et al., 2016)
CAS	sLCL, MS	Italy	1	ERS1100709	(Chiara et al., 2016)

TABLE 1.2: All published whole genome sequences to date (April 2016). BL: Burkitt's lymphoma; GC: gastric carcinoma; HL: Hodgkin's lymphoma; IM: infectious mononucleosis; LC: Lung carcinoma; LCL: lymphoblastoid cell line; sLCL: spontaneous lymphoblastoid cell line; MS: Multiple sclerosis; NPC: nasopharyngeal carcinoma; PTLD: posttransplant lymphoproliferative disease; PNG: Papua New Guinea; N. Africa: North Africa; *: Genomes assembled from read data of the 1000 Human Genomes Project; [1]: same donor.
Sequences included in the comparative genomics analysis in chapter 4 are marked with ● .

1.8 Outline

Studying variation of EBV can answer many interesting questions: It can be a means to better understand how virus variation impacts on pathogenesis of disease. It further bares the evolutionary history in relation to its host and distribution across the world.

This thesis centres around genomic studies on the impact of host/virus interaction in EBV infection using deep sequencing. I will present the results of these studies in three chapters, which will deal with the following topics:

1. The optimisation of a target enrichment method to sequence EBV directly from clinical material where human DNA is present in vast excess.

2. A comparative genomics analysis to elucidate the evolutionary history of EBV and the relationship between different world-wide isolates. This includes the identification of polymorphic genomic regions and of sites under selection as well as a comprehensive study of recombination through the generation of genome-wide linkage data. This data is used to identify potential targets of T cell immunity. Additionally, the population structure of EBV is examined.

3. The sequencing and analysis of whole EBV genomes from various clinical settings, including infectious mononucleosis and immunosuppression with chronic EBV infection, in terms of their inter- and intrahost diversity, as well as the comparison of genomes derived from Asian patients with carcinoma and primary infection.

Chapter 2

Materials & Methods

2.1 Materials & Reagents

2.1.1 Samples

Cell lines

DNA isolated from JSC-1, a (latent) EBV-infected, and Kaposi's Sarcoma Herpesvirus co-infected, primary effusion cell line (Cannon et al., 2000).

Clinical samples

Sample	Date	Source	Age	Sex	From	VL [copies/ml]	Diagnosis
ebv1	19/10/2011	Blood	15	f	GOSH	22,110	PTLD
ebv2	21/10/2011	Blood	15	f	GOSH	24,782	PTLD
ebv3	01/11/2011	Blood	15	f	GOSH	972	PTLD
ebv4	20/02/2012	Blood	15	f	GOSH	(CT) 34	PTLD
ebv5	01/03/2012	Blood	15	f	GOSH	6,078	PTLD
ebv6	09/05/2012	Blood	5	m	GOSH	942,609	Heart transplant
ebv7	08/05/2012	Blood	7	m	GOSH	444,984	Immunodeficiency
ebv8	18/05/2012	Blood	7	f	GOSH	1,200,000	Turner's syndrome, Heart transplant
ebv9	18/05/2012	Blood	3	f	GOSH	>2,000,000	Heart transplant
ebv11	23/05/2012	Blood	4	f	GOSH	6,950,932	Cystic fibrosis, Lung transplant
ebv13	15/04/2013	Blood	6	f	GOSH	> 2,000,000	Kidney transplant
ebv14	22/04/2013	Blood	5	m	GOSH	> 2,000,000	SOT
ebv15	31/05/2013	Blood	13	f	GOSH	1,295,240	SOT
ebv16	26/11/2014	Blood			GOSH	497,026	SOT
ebv17	26/03/2014	Blood			GOSH	1,067,180	SOT
ebv18	02/12/2014	Blood			GOSH	131,685	SOT
ebv19	27/03/2014	Blood			GOSH	714,159	SOT
ebv20	07/10/2014	Blood			GOSH	26,011	SOT
ebv21	06/01/2014	Blood			GOSH	2,525,480	SOT
ebv22	07/08/2014	Blood			GOSH	4,684,310	SOT
ebv23	06/08/2014	Blood			GOSH	1,458,050	SOT
ebv24	27/11/2014	Blood			GOSH	7,516	SOT
ebv25	15/01/2014	Blood			GOSH	14,532,600	SOT
ebv26	04/12/2014	Blood			GOSH	98,249	SOT
ebv27	19/02/2014	Blood			GOSH	838,767	SOT
ebv28		Blood			GOSH	10,298	SOT
ebv29		Blood			GOSH	11,914	SOT
ebv30		Blood			GOSH	473,352	SOT
ebv31		Blood			GOSH	1,130,850	SOT

Sample	Date	Source	Age	Sex	From	VL [copies/ml]	Diagnosis
P1-B1	26/05/2015	Blood			GOSH	66,473	PTLD
P1-T1	29/08/2015	Tumour			GOSH	11,228,500	PTLD
P2-B1	04/10/2012	Blood			GOSH	21,973	PTLD
P2-B2	18/11/2015	Blood			GOSH	1,707,130	PTLD
P2-B3	18/01/2016	Blood			GOSH	4,534,200	PTLD
P2-T1	02/10/2012	Tumour			GOSH	ND	PTLD
P3-B1	03/11/2011	Blood			GOSH	22,110	PTLD
P3-B2	03/09/2014	Blood			GOSH	952,152	PTLD
P3-T1	09/09/2011	Tumour			GOSH	ND	PTLD
P4-B1	15/03/2003	Blood			GOSH	>1,000,000	PTLD
P4-T1	26/03/2003	Tumour			GOSH	ND	PTLD
						VL [copies/μg]	
P1-812	01/06/2007	Blood	6	f	Toyoake, JP	114,182	IM
P1-833	12/06/2007	Blood	6	f	Toyoake, JP	238	IM
P2-1213	07/12/2007	Blood	7	m	Toyoake, JP	124,641	IM
P2-1246	19/12/2007	Blood	7	m	Toyoake, JP	13,973	IM
P3-2670	18/04/2009	Blood	4	m	Toyoake, JP	250,054	IM
P3-2740	11/05/2009	Blood	4	m	Toyoake, JP	17,173	IM
P4-2274	19/12/2008	Blood	6	f	Toyoake, JP	139,271	IM
P4-2392	06/02/2009	Blood	6	f	Toyoake, JP	3,809	IM
P5-1294	11/01/2008	Blood	15	f	Toyoake, JP	59,710	IM
P5-1323	18/01/2008	Blood	15	f	Toyoake, JP	6,964	IM
P6-1751	23/05/2008	Blood	3	m	Toyoake, JP	49,657	IM
P6-1789	04/06/2008	Blood	3	m	Toyoake, JP	346	IM
P7-2315	14/01/2009	Blood	11	f	Toyoake, JP	10,277	IM
P7-2634	03/04/2009	Blood	11	f	Toyoake, JP	24,275	IM
P8-414	13/10/2006	Blood	7	m	Toyoake, JP	22,197	IM
P8-516	15/12/2006	Blood	7	m	Toyoake, JP	8,982	IM
P9-2631	08/04/2009	Blood	4	m	Toyoake, JP	62,666	IM
P9-2645	13/04/2009	Blood	4	m	Toyoake, JP	37,832	IM
P10-2187	31/10/2008	Blood	1	m	Toyoake, JP	4,221	IM
P10-2777	20/05/2009	Blood	1	m	Toyoake, JP	212	IM
P11-871	01/07/2007	Blood	1	m	Toyoake, JP	24,101	IM
P11-920	24/07/2007	Blood	1	m	Toyoake, JP	3,487	IM
P12-1026	29/08/2007	Blood	11	m	Toyoake, JP	2,844	IM
P12-1078	25/09/2007	Blood	11	m	Toyoake, JP	1,142	IM

TABLE 2.1: All clinical samples processed including patient information where available. VL: viral load in copies/ml blood or copies/μg DNA, respectively. GOSH: Great Ormond Street Hospital for children, Great Ormond St, London, UK; Toyoake, JP: Department of Pediatrics, Fujita Health University School of Medicine, Toyoake, Japan; PTLD: post-transplant lymphoproliferative disorder; SOT: solid organ transplant; IM: infectious mononucleosis.

2.1.2 Reagents

Reagent	Supplier (Cat. No)
QIAamp DNA Blood Mini Kit	Qiagen (51106)
QIAamp DNA mini kit	Qiagen (51306)
Genomphi v3	GE Healthcare (25-6601-24)
Qubit dsDNA HS Assay Kit	Invitrogen (Q32854)
Genomic DNA Clean & Concentrator	Zymo Research (D4011)
Agencourt AMPure XP beads	Beckman & Coulter (A63881)
Illustra GenomiPhi V2	GE Healthcare (25-6600-30)
Qubit dsDNA HS Assay Kit	Thermo Fisher Scientific (Q32851)

TABLE 2.2: Reagents used in DNA preparation of clinical samples.

Reagent	Supplier (Cat. No)
SureSelectXT Reagent kit, MSQ	Agilent (G9612B)
Herculase II Fusion DNA Polymerase	Agilent (600677)
Dynabeads MyOne Streptavidin T1	Thermo Fisher Scientific (65602)
1X Low TE Buffer	Thermo Fisher Scientific 12090-015)
DNA 1000 Kit	Agilent (5067-1504)
High Sensitivity DNA Kit	Agilent (5067-4626)
D1000 ScreenTape	Agilent (5067-5582)
D1000 Reagents	Agilent (5067-5583)
High Sensitivity D1000 ScreenTape	Agilent (5067-5584)
High Sensitivity D1000 Reagents	Agilent (5067-5585)

TABLE 2.3: Reagents used for targeted enrichment.

Reagent	Supplier (Cat. No)	Assay
QuantiFast SYBR Green PCR kit	Qiagen (204754)	EBV load
QuantiTect SYBR Green PCR kit	Qiagen (204143)	Human load
Phusion	New England Biolabs (M0530)	PCR
Phusion HF buffer	New England Biolabs (B0518S)	PCR
GeneRuler 1kb DNA ladder	Thermo Fisher Scientific (SM0312)	Electrophoresis
Invitrogen 100 bp ladder	Thermo Fisher Scientific (15628019)	Electrophoresis

TABLE 2.4: Reagents used in PCR and qPCR.

Reagent	Supplier (Cat. No)
100 % Ethanol, molecular biology grade	Sigma-Aldrich (E7023)
Nuclease-free Water	Thermo Fisher Scientific (AM9932)

TABLE 2.5: Other general reagents.

Name	Orientation	Sequence	Fragment length	Start	End
PSE1	F	CCCAGAACAGCACCCGAAA		98700	98719
PSE4	R	GGGTAGATGGCGAGACTCTT	484	99164	99184
PSE5	R	GATAGACTGGGAGGCCTGA	254	98936	98955
PSE9	F	CTGGTTGATTACGGGGCACT		106976	106996
PSE10	R	GCCTCTGTCTCCTGGTTGAC	1000	107956	107976
PSE11	R	TTGTTGGAGACTACGTCCGC	1971	108927	108947

TABLE 2.6: Primers for shearing experiment (synthesised by Sigma-Aldrich). PSE1 was paired with PSE4 and -5, PSE9 was paired with PSE10 and PSE11 to produce a product of the fragment length listed. Start and end coordinates refer to the JSC-1 genome.

Name	Orientation	Sequence	Assay
KRAS_F	F	GCCTGCTGAAAATGACTGAATATAAAC	human load
KRAS_R	R	TGATTCTGAATTAGCTGTATCGTCAAG	human load
EBV_F	F	CCGGTGTGTTCGTATATGGAG	EBV load
EBV_R	R	GGGAGACGACTCAATGGTGTA	EBV load

TABLE 2.7: Primers for qPCR (synthesised by Sigma-Aldrich). The EBV primers are published in Wandinger et al., 2000.

2.2 Experimental methods

2.2.1 Whole genome sequencing of EBV from clinical samples

DNA extraction and preparation

Total DNA was extracted from whole blood using the QIAamp blood mini kit (Qiagen) and from frozen tumour samples using the QIAmp mini kit. DNA was quantified using the Qubit 2.0 Fluoromoeter and the dsDNA HS kit (Life Technologies). Whole genome amplification (WGA) was performed additionally for some clinical samples using Genomiphi V2 (GE Healthcare) according to manufacturer's instructions (table 2.2).

Targeted enrichment of EBV-specific sequences

For targeted enrichment, the SureSelect^{XT} protocol (further referred to only as SureSelect) was used. The protocol involves the following steps:

1. Shearing of genomic DNA samples
2. Preparation of SureSelect libraries
3. Amplification of SureSelect libraries
4. Hybridisation of DNA samples to RNA baits and capture
5. Amplification of captured libraries with indexing primers

For each sample, 3 µg or 200 ng of DNA was sheared using either the Covaris E210 or the E220 system. Illumina paired-end sequencing libraries were constructed using the standard SureSelect protocol, version 1.4.1 or version 1.6, respectively, with given additional alterations to optimise the protocol as described (see chapter 3). Before 2015, SureSelect was performed manually; from 2015 onwards, the Agilent Bravo Automated Liquid handling platform (henceforth referred to as the automation system) was used.

SureSelect baits Samples were enriched for EBV sequences as previously described (Depledge et al., 2011).

Two sets of RNA baits were used throughout the project. Both were designed by Daniel Depledge using the Agilent SureDesign software (Depledge et al., 2011). Set 1 was based on the published whole genomes sequences available in 2012, including published genomes from Baer et al., 1984; de Jesus, 2003; Dolan et al., 2006; Zeng et al., 2005; Liu et al., 2011; Kwok et al., 2012, while set 2 was created as an update in 2015 to include the diversity found in newly published genomes (extending it by the genomes published by Lin et al., 2012; Lei et al., 2013; Tsai et al., 2013; Kwok et al., 2014 and Palser et al., 2015). Both sets of baits were designed as overlapping 120-mers spanning the whole length of the positive strand of the genome with a 5x coverage for set 1 and 12x for set 2. Baits were synthesised by Agilent Biotechnologies.

Bait set 1 was used for all samples prepared up to 2015. This includes the optimisation experiments presented in chapter 3, and the Japanese infectious mononucleosis data set (on the automation system) which was then sequenced on a MiSeq (chapter 5). The bait set 2 was used again on the IM data set, the additional samples of immunocompromised children from the UK, as well as the paired tumour and blood samples (chapter 5).

Standard Agilent barcodes were added and all recommended quality control steps included (using either a Agilent 2100 Bioanalyzer or an Agilent 2200 TapeStation).

Next generation sequencing

Libraries were either sequenced on the Illumina MiSeq platform with a v3 reagent kit (300 cycles), or on the Illumina NextSeq with a 500/550 v2 reagent kit (300 cycles). Samples were demultiplexed using the automated workflow on the MiSeq to generate paired

FASTQ files. For the NextSeq generated reads, an in-house script was used for demultiplexing.

2.2.2 Human and viral load assay

To determine human and viral loads, qPCRs were performed using primers listed in table 2.7. If possible samples were prepared in duplicate. The gene *KRAS* was targeted for the human load, while *EBNA1* was targeted for the EBV load as previously described (Jabs et al., 2001) (table 2.4).

2.2.3 PCR to test shearing efficiency of episomal DNA

In order to test whether the shearing process works for episomal DNA, a PCR assay was used. Table 2.6 lists the primers and their coordinates as well as the expected fragment length.

The PCR was set up in a 25 µl reaction with Phusion polymerase according to the manufacturer's protocol (table 2.8a).

Reagent	[µl]
Nuclease-free water	15.75
HF Buffer	4
10 mM dNTPs	0.5
1 µM Forward Primer	1.25
1 µM Reverse Primer	1.25
Phusion	0.25
DNA	1

(A) Master mix.

Step	Temp [°C]	Time
Initial denaturation	98	30 s
35 cycles	98	20 s
	X	30 s
	72	X
Final extension	72	10 min
Hold	4	∞

(B) PCR cycler program; X specified in text.

TABLE 2.8: PCR conditions for shearing experiment with Phusion polymerase.

PSE1/4 and PSE1/5 were run with an annealing temperature of 62°C and an elongation time of 15 s. PSE9/10 and PSE9/11 were annealed at 64°C and 65°C with an elongation time of 30 s and 60 s, respectively.

PCR products of PSE9/10 and PSE9/11 were run on a 1 % and product of PSE1/4 and PSE1/5 on a 2 % agarose gel in TAE buffer.

2.3 Computational methods

2.3.1 Genome assembly

Different approaches were used to assemble the genomes of EBV throughout the work. Except JSC-1, all samples have been *de novo* assembled. There are two final *de novo*

pipelines (see below). Samples ebv6-15 have been assembled using CLC workbench, whereas all other samples were processed using the open source pipeline using SPAdes.

Reference-guided assembly

JSC-1 was first assembled using a reference-guided approach. Reads were trimmed with the FASTX toolkit (*FASTX toolkit*). Duplicates and low quality reads were removed using QUASR v7 (*QUASR*) using a minimum quality threshold of 30 and minimum read length of 50 bp.

Reads were then mapped against the type 1 reference genome with bwa 0.7.5a using the algorithms aln and sampe (Li and Durbin, 2009). SAM files were then further processed into BAM files using the samtools (Li et al., 2009) and a consensus generated using QUASR.

De novo assembly using CLC Genomics Workbench

This assembly pipeline is referred to as *pipeline 1* in the respective results chapters.

Paired-end sequence reads were processed using CLC Genomics Workbench 7 including the CLC Microbial Genome Finishing Module (Qiagen).

Prior to assembly, reads were pre-processed: Duplicated reads were removed, the remaining reads were trimmed based on base quality and further trimmed to a fixed length (50 bp at the 3′ end and 10 bp at the 5′ end). To assess the percentage of reads mapping to the target genome (on target reads, OTR) and to filter for EBV specific reads, all reads were mapped against all available EBV genomes (standard parameters). Non mapping reads were discarded, the remaining EBV specific reads were assembled using the CLCbio *de novo* algorithm with standard parameters (*CLC Workbench 7*). If the number of reads exceeded 200-300k, the data set was subsampled as this process was found to improve the assembly. The resulting contigs were then aligned against the reference genome NC_007605 and manually merged. If the data set had been subsampled, previously excluded reads were mapped back to the contigs. A consensus sequence for each genome was extracted under default parameters. Low coverage regions were defined as having a read depth of <20 (Illumina Inc., 2010) and sequences here were written as N.

This pipeline was mainly used for the initial genome assemblies in chapter 3 and later abandoned due to the proprietary character of the software. Moreover, the assembler, while giving good results, is very slow and additional processing requires many manual steps, in particular in the case of EBV due to the high number of contigs which likely arise because of the high number of repeat regions in the EBV genome.

De novo assembly using open source tools

This assembly pipeline is referred to as *pipeline 2* in the respective results chapters.

A mostly open source assembly pipeline was developed in our lab by Sofia Morfopoulou, and was used for the majority of the samples (chapter 5).

First, reads were trimmed using TrimGalore v0.3.7 (*TrimGalore*) with a quality threshold of 20 and then aligned with BLASTN (Altschul et al., 1990) against a database of 124 EBV type 1 and 2 genomes in order to extract EBV-specific reads. Those were then *de novo* assembled using SPAdes v3.5.0 (Bankevich et al., 2012) into contigs, which were filtered for those of greather length than 200 bp with quast v2.3 (Gurevich et al., 2013). The order and orientation of contigs were determined with BLASTN and a scaffold being generated using an in-house R script. The quality checked reads were re-mapped against this pseudogenome using BBMap (*BBMap*) and further processed into BAM-files with samtools v0.1.19 (Li et al., 2009) and Picard (*Picard*). The consensus was extracted with QUASR (*QUASR*), with the following parameters: minority bases were included as an ambiguity code at a frequency of 0.5, minimal base quality was set to 20 and minimal depth to 20 (Illumina Inc., 2010). All bases below depth 20 were written as N.

2.3.2 Sequence and phylogenetic analysis

Multiple sequence alignment

Multiple sequence alignments were obtained using mafft v7 (Katoh and Standley, 2013) and manually corrected using the alignment editor of MEGA (Tamura et al., 2013). SNPs were called based on differences to the reference genome WT (Accession No. NC_007605) and statistics calculated using the R packages **adegenet** (Jombart, 2008), **ape** (Paradis, Claude, and Strimmer, 2004) and **pegas** (Paradis, 2010).

EBV-typing

Determination whether a genome is of type 1 or 2 was done calculating the genetic distance under the Kimura-2-parameter model (K80) (Kimura, 1980) for the CDS sequences of *EBNA2*, *EBNA3A*, *EBNA3B* and *EBNA3C* to the type 1 and 2 reference strains (WT and AG876).

Nucleotide diversity

Nucleotide diversity is a measure of genetic variation and is defined as the average number of differences between two sequences. It is defined as

$$\pi = \sum_{i}^{n} \sum_{j}^{n} x_i x_j \pi_{ij}$$

where x_i and x_j are the respective frequencies of the ith and jth sequences, π_{ij} is the number of nucleotide differences per nucleotide site and n is the number of sequences (Nei, 1987).

Nucleotide diversity was calculated as implemented in **pegas** either for the whole ORF or in sliding windows of 100 bp with a step size of 1 bp for the whole genome alignment.

A SNP density plot was generated in **adegenet**. To test whether SNPs are randomly distributed across the genome, a Monte Carlo test as implemented in the same R package was run with 999 replicates.

Intrahost nucleotide diversity

Intrahost nucleotide diversity was computed for individual samples using an in-house program developed by Prof. Richard Goldstein and Juliana Cudini. It is defined as the average number of nucleotide differences between reads at a site (Nei and Li, 1979). Strand bias and random error rates were estimated and corrected using maximum likelihood methods.

Minority variants

In order to call minority variants (i.e. variants with a frequency less than 0.5) within one patient sample, quality checked reads were mapped against the consensus sequence generated during genome assembly with BBMap v35 (*BBMap*). In the case of assemblies generated with pipeline 1, raw reads were re-processed and mapped with BBMap against the original consensus. The sam-file was converted into a bam-file and sorted with samtools v0.1.19 (Li et al., 2009). Duplicates were removed using the `MarkDuplicates` tool of the Picard toolbox v1.138 (*Picard*). The bam-file was then converted into a pileup-file with samtools using the `mpileup2cns` option. All possible variants were called with VarScan v2.3.7 (Koboldt et al., 2012), without any prior filtering except for the minimal average mapping quality set to 20 (i.e. settings were as followed: `--min-avg-qual 20 --min-var-freq 0 --p-value 99e-02 --min-coverage 1`). Variants were then filtered using an in-house python program by Dr. Samit Kundu called varsnp v2.1, which allows to set a minimal frequency cut-off, a minimal coverage cut-off (see results chapter) and a minimal number of reads per strand (set to 2). The resulting variants were then checked for strand bias using an in-house R script that excludes variants where >90 % of bases were mapped to one strand. Additionally variants that were called in areas that mapped to repeat regions were excluded.

In some cases, mapping files have been randomly subsampled to a certain number of reads prior to variant calling. This was to achieve comparable read depth across samples, as read depth otherwise biases the number of minority variants called. The subsampling was done using the `DownsampleSam` tool of the Picard toolbox.

Principal component analysis

Principal components analysis (PCA) is a commonly used tool to find patterns in complex data, as it reduces the dimensionality of the data while retaining most of its variation. In genomics, this is for example applied to detect population structure, clusters, or potential outliers.

The aim is to convert a set of variables into a new set of linearly uncorrelated variables which are called the principal components (PCs). These PCs are linear combination of the

original variables and are ordered according to how much of the variation they contain. This allows the exploration of the data, e.g. in a two dimensional plot of two PCs, onto which the data is projected. Additionally, it enables the identification of original variables that contribute most to the variance captured by a PC.

PCA was performed by using the implementation for SNP data in the R package **adegenet** (Jombart, 2008) on the data sets further discussed in the individual chapters. Different combinations of PCs were explored in two-dimensional scatterplots. Variables (here, SNPs) contributing most to a certain PC were identified via their individual loadings, i.e. when their absolute loadings were higher than the third quartile of all absolute loadings.

A complementary Neighbor joining (NJ) tree was calculated with **ape**. Note that this is not representing a reliable phylogenetic tree, but is used to better visualise individual genomes and clusters from the PCA scatter plots.

Phylogenetic and recombination analysis

Short introduction to phylogenetic trees and networks Phylogenetic trees are graphical representations of the inferred evolutionary history of genes or organisms (taxa or operational taxonomic unit, OTU). The joining of two branches implies a common evolutionary origin. Various algorithms have been developed to infer a phylogenetic tree, from distance-based methods to Maximum Likelihood (ML) and Bayesian approaches.

Distance based methods calculate a distance metric which measure how dissimilar each pair of taxa are to each other. The simplest measure of distance is called *p-distance* and is the number of nucleotide differences per site. However, in the case of recurrent mutations, this model underestimates the true genetic distance, as it solely depends on the observed states in the data. For this reason, genetic distances are inferred under some evolutionary model which takes these scenarios into account.

The simplest model is the *Jukes-Cantor-Model*, also referred to as *JC69* (Jukes and Cantor, 1969). Here, all base frequencies are assumed to be equal (i.e. 0.25) as are the substitution rates between bases. Slightly more complex are Felsenstein's *F81* model (Felsenstein, 1981), which allows for varying base frequencies, and the *K80* model (Kimura, 1980), which takes into account different rates between transitions (A to G and C to T, i.e. purine to purine and pyrimidine to pyrimidine substitutions) and transversions (pyrimidine to purine or the other way around). The *HKY85* model combines the two latter models by allowing both for unequal base frequencies as well the distinction between transitions and transversions (Hasegawa, Kishino, and Yano, 1985). A further extension from here is the *TN93* model (Tamura and Nei, 1993), where in addition two cases of transitions (purine and pyrimidine transitions) are allowed to have varying rates. The most complex model, is the *GTR* (generalised time-reversible) model (Tavaré, 1986), where all parameters are free, i.e. varying rates for every kind of substitution, no matter whether it is a transversion or transition, as well as variable base frequencies.

One common distance based method to infer a phylogenetic tree is the Neighbor joining (NJ) algorithm. It assumes at the beginning a star-like topology of the tree without

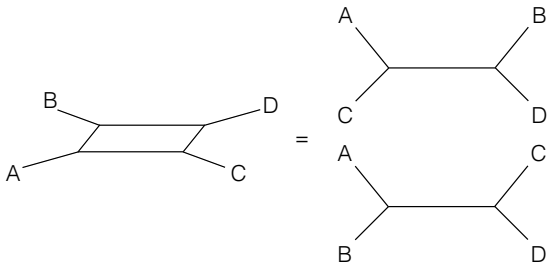


FIGURE 2.1: The split network on the left represents the two conflicting phylogenetic trees on the right. (After Lemey, P., Salemi, M., and Vandamme, 2009)

internal branches. Internal branches are introduced by joining taxa A and B together in such a way that the length of the internal branches (and by that the total length of the tree) is minimised. The length of the joining branch to A is dependent on the average distance of all taxa to A minus the average distance of all remaining taxon pairs.

Distance based methods usually result only in one tree. Maximum Likelihood methods, on the other hand, calculate different tree topologies and try to find the optimal or best-fitting tree. As the names states, their optimality criterion aims to find the most likely tree, i.e. it considers the probability that a certain tree topology could have given rise to the observed data given an evolutionary model. There are different heuristic search algorithms to find the tree, but they are all generally quite computationally expensive. However, they deliver a statistical framework to evaluate the support of branches.

Phylogenetic trees make the assumption that the underlying evolutionary process can be explained as a linear, tree-like process (Huson, 1998). In the case of recombination, however, a number of different tree topologies might be suitable to support the data, as recombination allows the exchange of whole fragments between taxa (for example different virus isolates of the same viral species). Phylogenetic networks are therefore a better way of representing this kind of data. However, their conception can be quite varied, but are defined by Huson and Bryant, 2006 as *any* network in which taxa are represented by nodes and their evolutionary relationship by edges. Thus, it includes phylogenetic trees as a special case. Other types can be *split networks*, which depict incompatible or ambiguous signals within the data through parallel edges. Internal nodes do therefore not necessarily represent ancestral species. Hence, these networks can be seen as "implicit" representation. In *reticulate networks*, on the other hand, additional edges connecting "conventional" tree branches describe a putative evolutionary history, i.e. internal nodes do represent ancestral species. This makes these networks an "explicit" phylogenetic representation (Huson and Bryant, 2006).

Construction of phylogenetic trees If not stated otherwise, phylogenetic trees were generated using the program *raxml*, which uses a Maximum Likelihood approach under the GTR model of nucleotide substitution and the Gamma model of rate heterogeneity with a rapid bootstrap analysis of 500 runs.

Split network Split networks were generated in Splitstree4 using the uncorrected p-distance and the NeighborNet algorithm (Huson and Bryant, 2006). It represents a set of incompatible splits from the data.

Testing for presence of recombination A test for the presence of recombination is the PHI-test, which is simple and robust. It is able to distinguish between recurrent mutations and recombination in various circumstances (Bruen, Philippe, and Bryant, 2006). The test statistic Φ_w (or PHI, pairwise homoplasy index) is based on the notion of compatability/incompatability between sites. A pair of sites is compatible if their genealogical history can be explained without involving homoplasies (convergent or recurrent mutations) or recombination events. Φ_w is based on the calculation of the a mean incompatibility score between nearby sites up to w bases apart, and can be interpreted as the minimum number of homoplasies that are needed in any tree to explain the history of the pair of sites (Bruen, Philippe, and Bryant, 2006).

The test was both employed on the whole genome alignment of 83 type 1 sequences, various subsets of SNPs, see Chapter 4, as well as in a sliding window approach across the whole genome alignment in windows of 100 bp with PhiPack (Bruen, Philippe, and Bryant, 2006).

Determination of recombination breakpoints Recombination may bias the estimation of dN/dS (i.e. the ratio of synonymous to nonsynonymous substitution rates) as a measure for molecular adaptation at the codon level by increasing the number of false positively selected sites (Anisimova, Nielsen, and Yang, 2003). It is therefore important to take into account the effect of recombination prior to conducting selection analysis. This can be done by determining the recombinant fragments, for which then individual phylogenetic trees are reconstructed. Using the relevant tree for each fragment, dN/dS ratios are then estimated separately for every region (see section "Selection analysis") (Pérez-Losada et al., 2015).

To detect possible recombination breakpoints, the GARD algorithm (Kosakovsky Pond et al., 2006) which is part of the HyPhy v.2.2 package (Pond, Frost, and Muse, 2005) was used.

GARD (Genetic Algorithm Recombination Detection) works by identifying whether there is phylogenetic incongruence in a multiple sequence alignment and allows thereby the detection of more than one recombination breakpoint. It splits the alignment into a growing number of non-recombining fragments, reconstructs a phylogenetic tree for every fragment and uses then the small sample Akaike's Information Criterion to evaluate the goodness-of-fit and choose the best model. The best breakpoints are determined either exhaustively (for two fragments) or heuristically (for more than two) (Kosakovsky Pond et al., 2006).

The coding sequences (CDS) of all ORFs (repeats excluded) were scanned for recombination breakpoints prior to selection analysis. In cases of recombination, alignments were split at the estimated breakpoints under consideration of the reading frame.

Analysis of linkage disequilibrium

Linkage disequilibrium (LD) is the non-random association of alleles. It can be influenced by various factors, such as selection, the physical linkage between sites and recombination. Because recombination is more likely to occur between two distant sites, there is a negative relationship between LD between two sites and their physical distance between them. In other words, recombination reduces the LD between two sites.

Different measures for LD exist. A common measure is the the square of the correlation coefficient between the occurrence of different nucleotides at different sites. It is defined as

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$$

with

$$D = p_{AB}p_{ab} - p_{aB}p_{Ab}.$$

Conventionally, p_A is the frequency of nucleotide A at the first site and p_B the frequency of B at the second site etc. p_{AB} is the allele frequency with nucleotide A present at the first site and B present at the second (Balding, Bishop, and Cannings, 2007; Haydon, Bastos, and Awadalla, 2004). Lower case a and b signify the second nucleotide present at the respective sites.

Significance of the association can be assessed using the relationship between the test statistic of the contingency table test, X^2, and r^2

$$X^2 = n\hat{r}^2$$

where X^2 is asymptotically χ^2 distributed with one degree of freedom. However, for small samples sizes n, the χ^2 is unlikely to hold (Balding, Bishop, and Cannings, 2007), which is why in the case of our data set we chose a different test for LD.

Across the whole genome The whole genome alignment of type 1 sequences was restricted to its 5,190 biallelic sites where maximally one sequence was missing. LD was analysed between all possible combinations of biallelic SNPs. The significance of LD was assessed using Fisher's Exact test on the 2×2 contingency table, with a pair of SNPs being significantly associated if $p < 0.05$ under Bonferroni correction.

Detection of local clusters of LD In order to detect local signatures of linkage, a sliding window approach was used. Here, the p-values of a window were compared to the distribution of p-values drawn from the genome-wide set of comparisons with a Mann-Whitney-Wilcoxon U test. This "null distribution" of p-values was created by sampling the diagonal ribbon of the genome-wide association matrix with a quantile-function in order to create a representative sample of the same size as the window.

Using windows of varying size based on a fixed number of biallelic SNPs is problematic because the masked repeat regions increased the window sizes dramatically. Additionally, we found a correlation between the biallelic SNP density and the power of the test (Pearson's correlation test with r=0.14 and p = 1.06e-05). In consequence, windows of a fixed size of 1,400 bp were used in which 20 biallelic SNPs were uniformly sampled. This size is a trade-off between having enough statistical power across the genome and resolution (figure 2.2).

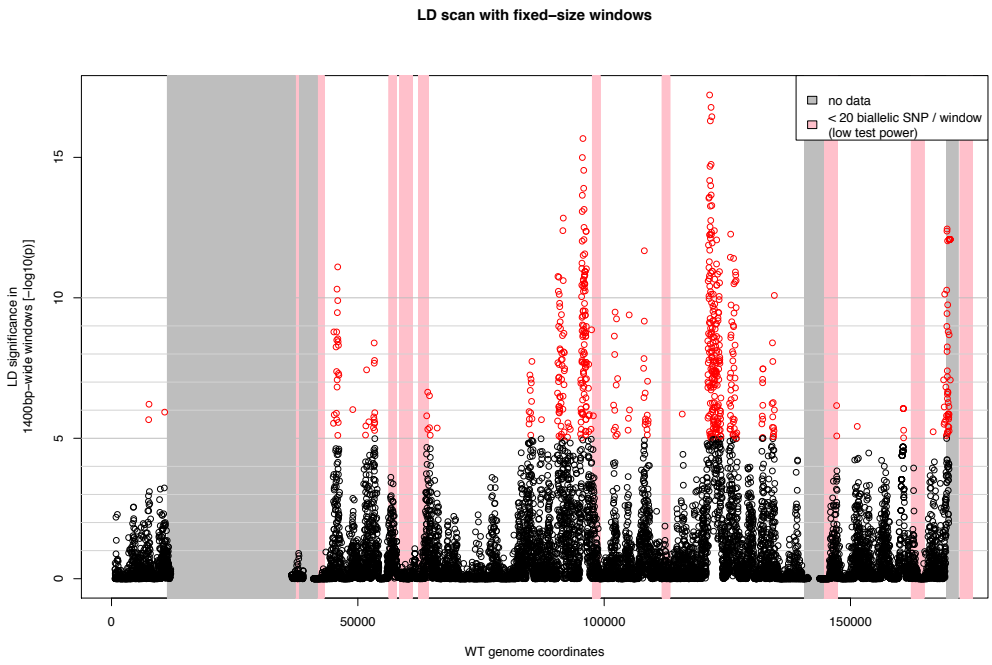


FIGURE 2.2: Genome map of linkage disequilibrium. LD score, computed in 1400 bp windows across the genome (WT reference coordinates). Values above 5 (circa top 5% values) are highlighted in red. The background is grey for areas of no data, and pink for polymorphism density of less than 20 biallelic SNPs per 1400 bp window, indicating a drastically reduced test power.

2.3.3 Gene network analysis

Short introduction to graph theory In order to understand the linkage between SNPs better, one can represent this as a network which can in turn be understood as graphs.

A graph G is defined as a pair (V, E) where V is a set of vertices (or nodes) and E is a set of edges which connect nodes, defined as $E = \{(i, j)|i, j \in V\}$. When two nodes i and j are connected with an edge, they are called *neighbours*.

Graphs can be either *undirected* or *directed*. In the latter case, the graph G is defined as an ordered triplet $G = (V, E, f)$ with f being a function that maps each edge in E to a ordered pair of vertices in V. In this work, however, we focus on undirected graphs.

A graph G can also be weighted. In that case, the set of edges E is associated with a weight function $w : E \to \mathbb{R}$. This can be interpreted as the relevance of a connection between two nodes.

Network of linked genes A network was constructed and analysed with the R package **igraph** (Csardi and Nepusz, 2006), where each node represents an ORF. An edge between two nodes was drawn if there was at least one pair of *nonsynonymous* SNPs in LD between them. The edges were weighted by a linkage score based on the number of linked nonsynonymous SNPs between two ORFs normalized for the ORF and genome lengths. The $n \times n$ matrix where n is the number of nodes containing all linkage scores between nodes i and j if there is an edge and 0 if there is no edge is called the adjacency matrix A.

The nodes of the network were divided into two sets of nodes, a) those ORFs known to encode antigens (immunogenic, IG) and b) those that do not (non-immunogenic, NIG). This classification was based on the experimentally confirmed antigens found in the IEDB database (*Immune Epitope Database*), with restriction to those antigens whose epitopes have been confirmed by at least two studies.

In order to identify important subgraphs/nodes, several methods were used:

Hierarchical clustering was performed using the distance matrix $D = 1 - A$, where A is the adjacency matrix (linkage score).

Another concept to rank nodes and thereby identifying important nodes in a network is *centrality*. There are two main notions how to interpret "importance". One notion relates to the flow or transfer across the network (not further considered here), the other one relates to the involvement of nodes in the cohesiveness of the network. There are many different centralities defined on different properties. *Degree centrality* is a ranking nodes based on their degree, *closeness centrality* ranks based on the average shortest path from a certain node to all other nodes and *betweenness centrality* ranks those nodes higher, that act as a bridge along the shortest paths between two other nodes.

Instead we chose *eigenvector centrality*, which measures the influence of a node in a network. In contrast to ranking nodes based on their degree, a node can have a high eigenvector centrality even if it is not highly connected. Instead, this node might have few, but *important* neighbours. The basic idea is therefore that influential nodes are influential, because they are connected to other influential nodes. (This notion is related to Google's PageRank algorithm.)

Given an adjacency matrix $A = (a_{i,j})$ of a network, the eigenvector centrality x_i of node i with k neighbours is defined as

$$x_i = \frac{1}{\lambda} \sum_k a_{k,i} x_k$$

where $\lambda \neq 0$ is a constant. This means, the centrality of the node i is the ith component of the first eigenvector x of A and the relation can be written as $Ax = \lambda x$. Because all

entries in the eigenvector are required to be non-negative, only the largest eigenvalue λ_{max} fulfills this (Pavlopoulos et al., 2011).

2.3.4 Population structure analysis

The population structure was analysed with *structure* v2.3.4 (Pritchard, Stephens, and Donnelly, 2000) using the automisation and parallelisation pipeline *strauto* v1.0 (Chhatre and Emerson, 2016).

Structure is a clustering method to infer population structure based on multilocus genotype data. It assumes the existence of k populations, to which individuals are probabilistically assigned. It also allows for individuals to be of admixed genetic background, in which case they are assigned to more than one population. The assumption of the model is that *within* populations loci are unlinked and in Hardy-Weinberg equilibrium.

Structure was run ten times for every k ranging from 1 to 10 with the admixture model and the correlated allele frequency model. The outputs are cluster member coefficient matrices, which contain the admixture proportions for each individual. To infer the number k clusters that best fits the data, the likelihood values of the models were assessed using Evanno's method (Evanno, Regnaut, and Goudet, 2005) as implemented in *structure harvester* (Earl and VonHoldt, 2012). This *ad hoc* approach estimates the number of clusters k using the statistic Δk, which is based on the rate of change in the log probability of the data between successive k values.

Because *structure* is based on stochastic simulation, every replicate run can result in different outcomes. Cluster labelling is random, i.e. what is called "cluster A" in replicate 1 can be "cluster C" in another replicate. Results for the best fitting k were therefore further analysed with CLUMPP (Jakobsson and Rosenberg, 2007) using the FullSearch algorithm, which permutates the cluster member coefficients of the result matrix of every replicate run such that replicate runs match each other as closely as possible. It also outputs the mean of the permuted matrices across replicates. The program *distruct* was then used to visualise results (Rosenberg, 2004).

2.3.5 Selection analysis

Considering the results from the recombination analysis, genes were then tested for the presence of detectable selection. Two tests were employed: Tajima's D, using the R package pegas, and codeml from the paml package v4.8 (Yang, 2007).

Tajima's D Tajima's D is a summary statistic method calculated over the whole gene, whose test statistic considers the relative frequency of polymorphic sites (Tajima, 1989). The observed values from an alignment are compared to the expected values assuming a null model of neutral evolution. This null model makes a number of assumptions: all mutations are either lethal or neutral, a constant effective population size and mutation rate, no recombination or migration, random mating and an infinite sites model, i.e. each mutation occurs at a different site. The test statistic D is defined as follows

$$D = \frac{\theta_\Pi - \theta_S}{\sqrt{Var(\theta_\Pi - \theta_S)}}$$

where θ_Π is the average number of pairwise differences (reflecting individual sequence divergence) and θ_S the number of segregating sites (reflecting population diversity). Both parameters are estimators for the population genetics parameter θ. If all assumptions of the null model of neutral evolution are met, they are expected to result from the same underlying processes (genetic drift) and their difference and in consequence D is zero. Selection, however, (and other violations of the null model assumptions) affects the values of the respective θ estimators, causing D to be non-zero. Depending on the direction of deviation from zero, different inferences can be made about evolutionary forces acting on a gene. Significance of D was assessed at a threshold of $\alpha = 0.05$ under assumption of a beta distribution (Tajima, 1989).

Positively selected sites with codeml The second selection test employed detects specific sites within a gene that are under positive selection (positively selected sites, PSS) using a Maximum Likelihood approach using the program codeml (Yang, 2007). A measure of selection is dN/dS, the ratio non-synonymous to synonymous substitution rates, also termed ω. Values <1, =1 and >1 indicate purifying selection, neutral evolution and positive selection, respectively.

Model	p	Parameters
M0 (one ratio)	1	ω
M1a (neutral)	2	$p_0, (p_1 = 1 - p_0)$, $\omega_0 < 1, \omega_1 = 1$
M2a (selection)	4	$p_0, p_1, (p_2 = 1 - p_0 - p_1)$, $\omega_o < 1, \omega_1 = 1, \omega_2 > 1$
M7 (beta)	2	p, q
M8 (beta & ω)	4	$p_0, (p_1 = 1 - p_0)$, $q, p, \omega_S > 1$

TABLE 2.9: Site-wise codon models used from the codeml program. p is the number of free parameters in the ω distribution (parameters in brackets are not free).

Input for this program are the codon corrected CDS alignments (or recombinant fragments) and their respective ML trees. Branch lengths of nucleotide trees represent estimated nucleotide substitutions per nucleotide site. To estimate the branch lengths as the number of nucleotide substitutions per codon site, model M0 was applied first.

Using site-wise models that allow ω to vary among codons, four models were tested in two pairs against each other in a likelihood-ratio test (LRT) (Anisimova, Bielawski, and Yang, 2002; Yang and Swanson, 2002): The M1a-M2a comparison tests a model of nearly neutral evolution (ω either <1 or =1) against a model of positive selection (ω can be <1, =1 and >1); The M7-M8 comparison tests two models where ω is assumed to follow

a beta distribution, which is bounded by 0 and 1, with M8 adding an additional class of sites with a free ω parameter, allowing for positive selection. The alternative model (M2a or M8) was considered to be a better fit with the Likelihood ratio test (LRT) with a significance threshold of $\alpha = 0.05$. Given a significant model for positive selection, the Bayes empirical Bayes (BEB) criterion was used to calculate the posterior probability of each site to be under positive selection. Hence, sites with $P > 95\ \%$ were considered to be PSS.

2.3.6 T cell epitope prediction

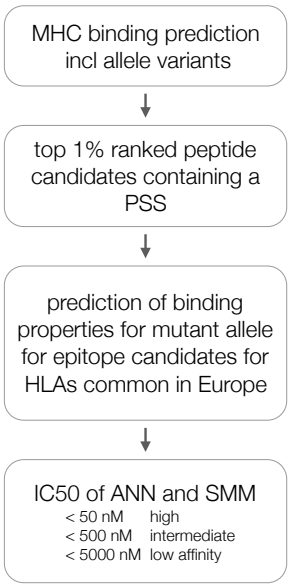


FIGURE 2.3: Scheme for epitope prediction.

T cell epitopes are peptides, usually derived from pathogens, that are presented on major histocompatibility complex (MHC) molecules and elicit a T cell mediated immune response. In humans, the MHC is also called human leukocyte antigen (HLA).

The IEDB analysis resource consensus tool was used to predict MHC binding for MHC class II (Wang et al., 2008; Wang et al., 2010a) and MHC class I (Kim et al., 2012) for selected EBV proteins that contain PSS. The consensus tool combines predictions from a number of algorithms (ANN (Nielsen et al., 2003; Lundegaard et al., 2008), SMM (Peters and Sette, 2005) and comblib (Sidney et al., 2008)). For the prediction, the HLA allele reference set was used which comprises the most common HLA alleles and is representative of commonly shared binding specificities (Vita et al., 2014).

Protein sequences including variant alleles were used as input for the prediction. The top 1% peptides from the MHC binding prediction were screened whether they contained one of the previously determined sites under positive selection. If so, the effect of observed variations of the PSS were compared regarding their binding properties as reflected by the estimated IC50 value to HLA alleles common in Europe. Peptides were considered to be putative high affinity epitope candidates if there was a difference between variants. A rule of thumb as given by IEDB is that IC50 values <50 nM are high,

<500 nM intermediate and <5000 nM low affinity (*Immune Epitope Database*; Sette et al., 1994).

Chapter 3

The application of target enriched whole genome sequencing of EBV to clinical blood samples

3.1 Introduction

Whole genome sequencing (WGS) of viruses allows the study of genetic variants that have an impact on disease development. It enables to assessment of intra- and interhost population structures, low-level drug resistance, and transmission. Moreover, WGS enables the identification of sites that are under selection, e.g. through host interactions. On a greater scale, WGS gives a comprehensive picture of variation across the genome, and allows study of populations on a local and global scale, and gives insight into the evolution of the viruses within patients.

Sequencing directly from clinical samples, however, is challenging as host DNA is present in vast excess to viral DNA. Direct sequencing of clinical samples in which pathogens are present leads to proportional representation of sequence reads derived from virus and host, but is only feasible in situations where virus is naturally present at extremely high titres, e.g. tumours in the case of EBV (Hsieh et al., 2007). This approach often requires large quantities of clinical material for successful virus genome recovery (Liu et al., 2011).

For this reason, several standard techniques are applied to enrich for the virus prior to sequencing, including *in vitro* culturing or amplification via polymerase chain reaction (PCR). Both methods can introduce novel polymorphisms that may bias results. PCR in particular has a higher error rate in low titre samples as more PCR cycles are needed to generate sufficient material for sequencing. For example, the high fidelity polymerase Phusion has a low error rate of 4.4×10^{-7} to 9.5×10^{-7} (compared to e.g. *Taq* polymerase with 2.28×10^{-5}), but this still leads to 1.32 to 2.85 % of DNA molecules containing an error after 30 cycles for a 1 kb template (according to manufacturer's data). In a comparative study of genome amplification methods, PCR has been found to introduce the highest bias (Pinard et al., 2006). Additionally, repeat regions, secondary structures and locally high GC-content can be PCR-inhibitory. Large genomes, such as those from

herpesviruses, are particularly problematic. Although technically possible, a large number of overlapping PCRs are required to span the whole genome and additional primer sets are needed in case of diverse genotypes or general high diversity. Recombination of co-amplified DNA molecules that are genetically distinct is another known problem, leading to artificial recombinants (Liu et al., 2014). Moreover, detection of minority variants can be hampered by uneven amplification across the genome (Houldcroft, Beale, and Breuer, 2017).

In the case of EBV, establishing a lymphpoblastoid cell line (LCL) may introduce a bias towards viral strains that most effectively immortalise B cells, thereby losing information about intrahost variation. Additionally, passaging may introduce other new genetic variations in the form of point mutations, deletions or genomic rearrangements. This has been described in CMV, for example (Dargan et al., 2010), but also the first sequenced genome of EBV, B95-8 (Baer et al., 1984) which was cultured in a marmoset B cell line and exhibited a large non-canonical deletion spanning the BamHI A fragment (de Jesus, 2003).

Targeted enrichment is a novel technique originating from exome sequencing, where specific regions of interest are selected for prior to deep-sequencing. This method was first adapted and applied to viral genome sequencing by Depledge et al., 2011 and has since been used for a number of distinct DNA and RNA viruses (Kwok et al., 2015; Palser et al., 2015; Depledge et al., 2014; Christiansen et al., 2017; Hage et al., 2017) as well as other pathogens (Melnikov et al., 2011; Brown et al., 2015; Christiansen et al., 2014) to sequence directly from clinical samples. This method has the advantage of using relatively few PCR cycles, the preservation of minor variant frequencies as seen *in vivo* (Depledge et al., 2014) and the possibility of automation (Depledge et al., 2011). However, costs for this method are high and it is by design limited to pathogens with known genomes.

Here, I will demonstrate the utility of target enrichment on clinical EBV samples by comparing and contrasting multiple protocols as well as the manual versus the automated preparation of samples. The focus will be on blood samples, which represent both the serum and lymphocyte compartment of EBV infection. These samples are from paediatric patients from two settings: immunocompromised transplant recipients and infectious mononucleosis. Both patient groups have elevated EBV titres in the blood compared to asymptomatic, healthy carriers, and are therefore ideal candidates to adapt target enrichment of EBV for WGS from blood. Sequencing directly from blood eliminates the need of culturing EBV prior to sequencing while also allowing the study of EBV from clinical settings that are non-malignant.

3.2 Results

3.2.1 Shearing efficiency of episomal DNA

The first step in the SureSelect protocol is the shearing of the DNA sample via ultrasonication, with a target fragment size of 150 to 200 bp. SureSelect was originally developed for exome sequencing, i.e. of DNA from long, linear chromosomes. The episomal nature of EBV DNA could potentially lead to less efficient shearing. While fragment size distribution is checked as a standard quality control step after shearing with a bioanalyzer, the human DNA in excess might mask if the viral DNA is not sheared properly. A simple PCR-based assay was therefore designed to test this. Primers were designed for the EBV episome using two forward primer and four reverse primers generating fragments in approx. 250 bp, 500 bp, 1 kb, and 2 kb size. A PCR was performed on sheared and non-sheared DNA samples of the EBV-infected cell line JSC-1.

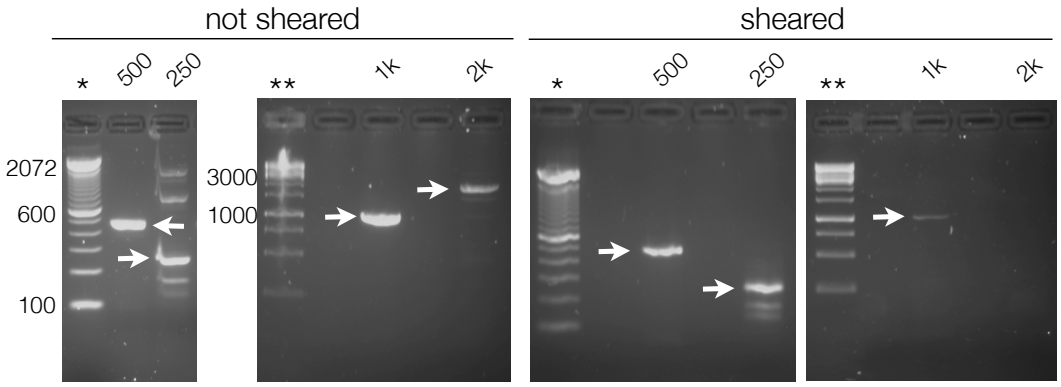


FIGURE 3.1: Shearing experiment. *: 100 bp ladder; **: 1 kb ladder.

Figure 3.1 shows the result of this PCR. All fragments could be detected in the non-sheared DNA samples, as expected (marked with an arrow). In the sheared DNA sample, the small fragments of 200 and 500 bp length could still be detected. The 1 kb fragment band was only faint, indicating that less of the product has been amplified, while the 2 kb product could not be detected at all. This indicates that larger fragments of the episome are either only present in smaller amounts or absent in the sheared sample, and confirms the typical fragment size distribution on a bioanalyzer chip (figure 3.2): a peak around 200 bp, a tail around 500 bp, and barely any signal for larger fragments.

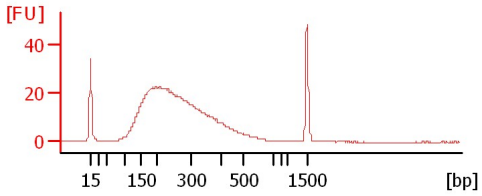


FIGURE 3.2: Typical fragment size distribution after shearing of genomic DNA.

This shows that episomal DNA is as efficiently sheared as long, linear DNA molecules.

3.2.2 Targeted enrichment of a EBV$^+$ cell line (JSC-1)

In order to verify and test the bait set used to select for EBV-specific reads, DNA from the cell line JSC-1 was enriched for EBV sequences using the manual SureSelect protocol with 3 μg of input DNA and sequenced on an Illumina MiSeq. Reads were quality controlled and assembled against the type 1 reference genome (NC_007605, here referred to as WT).

In total, 979,674 paired-end reads were obtained of which 656,700 were EBV-specific. This makes an on target read (OTR) percentage of 67 %, showing that enrichment worked very well. 97.3 % of the genome was covered with depth >20x with a mean depth of approximately 1000. Figure 3.3 shows the coverage plot, depicting the number of reads mapped across the genome.

Coverage is uneven across the genome. Many areas of low coverage correspond to the major repeat regions (shaded areas in the coverage plot). These areas cannot be confidently assembled or mapped given the short read lengths of Illumina sequencing. Reads are often shorter than the genomic repeats, meaning these short reads can map at multiple locations. Thus in all analyses of EBV genomes, these areas will be masked for this reason.

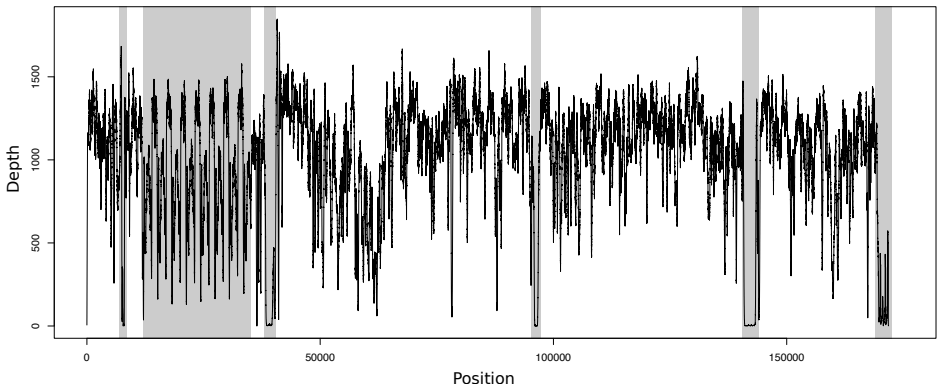


FIGURE 3.3: Number of mapped reads (coverage) against the WT reference of JSC-1 reads. Areas of low coverage often correspond to repeat regions shaded in grey.

3.2.3 Targeted enrichment of EBV from blood extracts

The targeted enrichment approach was applied to a first batch of different DNA extracts from blood of one immunocompromised, paediatric transplant patient from the UK. Prior to enrichment, samples were subjected to whole genome amplification (WGA). The enrichment protocol was the manual protocol with 3 μg input of DNA.

Table 3.1 lists the samples, viral loads and respective read counts as well as OTRs. It was not possible to retrieve full genomes, as all samples had <1000 OTR, which amounted to only 0.01 % to 0.03 % of EBV-specific paired-end reads, suggesting that enrichment failed.

While viral titres were not low in clinical terms, the titres were too low for successful enrichment as effectively only 30 to 1700 viral copies would be contained in 3 μg input material. Therefore, six additional DNA blood extracts from six immunocompromised

Sample	Viral load [copies/ml]	Read pairs	OTR pairs	OTR %
ebv1	22,000	1,110,989	259	0.02
ebv2	25,000	1,887,738	332	0.02
ebv3	1000	2,145,258	734	0.03
ebv4	34*	2,553,556	342	0.01
ebv5	6000	2,012,518	229	0.01

TABLE 3.1: Sample and sequencing information for blood DNA samples. Read pairs: total number of read pairs from the sequencer. OTR: on target reads, the number of paired-end reads mapped to EBV, without duplicate reads; *: CT value of the PCR product

children with a higher viral load were chosen. They were treated the same (manual 3 µg protocol). Table 3.2 lists the sample and sequencing data.

Sample	Viral load [copies/ml]	Read pairs	OTR pairs	OTR %
ebv6	943,000	805,280	956	0.12
ebv7	445,000	739,987	1,768	0.24
ebv8	1,200,000	903,760	4,885	0.54
ebv9	>2,000,000	1,047,922	14,661	1.40
ebv10	1,400,000	821,088	4,406	0.54
ebv11	710,000	1,356,031	5,765	0.43

TABLE 3.2: Sample and sequencing information for blood DNA samples with higher viral loads.

Viral loads for all samples were one to two orders of magnitude higher and enrichment improved as OTR percentages increased by one to two orders of magnitude. The highest OTR % was found in sample ebv9 with 1.4 % of paired-end reads mapping to EBV, which might be reflective of the extremely high viral load (only a minimum value of greater than 2 million copies/ml is available for this sample). However, compared to other herpesviruses previously sequenced in our group using SureSelect, OTR percentage was still far lower relative to viral load (personal communication with J. Breuer). For sequencing VZV from blood, for example, using the same 3 µg protocol, OTR percentage was 71 % (Depledge et al., 2011).

Varying amount of DNA input

As mentioned before, SureSelect was originally designed for exome sequencing. The application of this method to virus WGS, however, is quite different as non-target DNA is present in such vast excess. It was therefore investigated, whether the input in terms of total number of viral copies is more important for enrichment success than the input in terms of total amount of DNA.

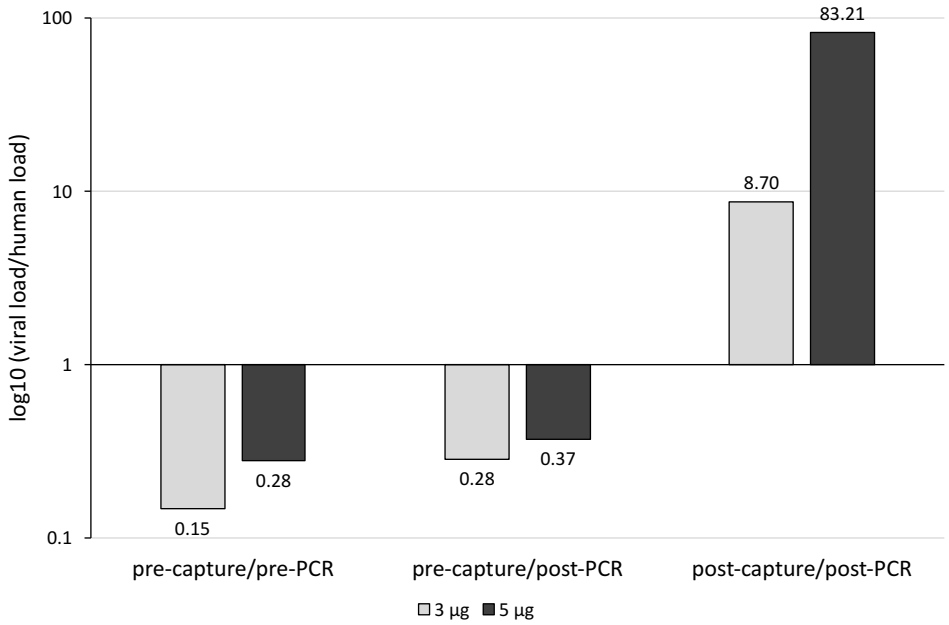


FIGURE 3.4: Ratio of viral versus human DNA at different steps in the SureSelect protocol on a logarithmic scale. Values at the end of the bars represent the actual (non-logarithmic) ratios.

Sample ebv9 was therefore treated with the 3 µg SureSelect protocol using either 3 µg and 5 µg total input DNA. At every important step of the protocol viral and human loads were determined by qPCR to identify steps in which viral copies might be potentially lost and to assess success of enrichment. These steps are: pre-capture (i.e. prior to hybridisation for enrichment) (1) before and (2) after the PCR step, and (3) post-capture after the final PCR amplification.

Figure 3.4 shows ratio of viral to human DNA at the three steps during the SureSelect protocol. Pre-capture, human DNA is present in excess compared to viral DNA (3 to 7-fold more human copies than viral ones). Here, the pre-capture PCR step decreases the difference between viral to human DNA, and a greater decrease can be observed in the sample using 3 µg total DNA input. Enrichment reversed the ratio in both samples. Using around 1.6 times more total DNA input resulted in a 10-fold enrichment increase. That is more than would be expected and could be due to a PCR bias leading to an increase in duplicate reads.

DNA input	Read pairs	OTR pairs	OTR %
3 µg	225,627	20,786	9.21
5 µg	167,964	18,635	11.09

TABLE 3.3: Sample and sequencing information for ebv9 with varying input amounts of DNA into the SureSelect protocol.

The final libraries were sequenced on a MiSeq. Table 3.3 shows the sequencing data.

In accordance with the qPCR results, using more input DNA resulted in a higher percentage of reads mapping to EBV, but the difference was not as stark as in the qPCR, as only approximately 1.2 % more reads mapped to EBV. This means that using 1.6 times more DNA input (and consequently also viral copies) resulted in an approximate 1.2-fold increase of OTR percentage. This suggests that using more total DNA input than the recommended protocol is not detrimental, and that the total number of viral copies in the input is an important factor for enrichment success.

Effect of WGA and controlling for loss

As clinical material is often scarce, a standard method for amplification of genomic material is WGA with Phi29 DNA polymerase, where random primers are used in a non-PCR reaction. I therefore wanted to test the effect of WGA with SureSelect on success of enrichment.

An additional source of problems for enrichment is loss of genetic material. The first part of the protocol (library preparation) contains numerous washing steps; at each of them, around 20 % of material is estimated to be lost (according to manufacturer's information). Because the number of viral genomes is a determining factor for success of capture, it was tested whether more viral copies are being lost during library preparation than would be expected.

To test this with a larger number of samples, the qPCR experiment from the section above was repeated for the remaining samples (ebv6, -7, -8, -11; there was not enough material left for ebv10) with fresh extracts of whole blood from the same patients. Due to scarcity of clinical material, the standard amount of 3 μg of input DNA was used. The qPCR was performed in duplicates for every sample collected at each of the three major protocol steps (as described above).

Previously, these samples were subject to WGA (table 3.2). Here, all four samples were used directly without WGA. While a comparison of the effect of WGA with the previous run would be possible, in order to account for preparation variability (see below), two of the samples were additionally treated with WGA in this experiment (ebv8G, ebv11G).

Viral loads were also measured for the fresh extracts. Table 3.4 shows the number of EBV genome copies that went into the SureSelect protocol. To account for the loss in the library preparation, which consists of four clean up steps, actual and expected total copy numbers were calculated, assuming that 20% are being lost at each of the four clean-up steps. For ebv6, the deviation from the expected value was greater with one order of magnitude, although it still fell within the range of one standard deviation (SD) of the actual number. But this SD of the actual copy number is fairly large due to a higher SD of the measured Ct value (>0.3), suggesting the actual copy number is not very accurate. For all other samples, the actual number of copies fell within the range of the expected number of copies (and all samples had an acceptable SD of the Ct value). This highlights, that the clean up steps are critical, but also suggests that they are not necessarily biased towards a more preferential loss of viral material.

Sample	# copies in 3 µg	exp. # copies	# copies	SD
ebv6	34,000	16,000	8,000	10,000
ebv7	32,000	13,000	19,000	11,000
ebv8	40,000	16,000	15,000	6,000
ebv11	31,000	13,000	11,000	2,000

TABLE 3.4: Expected and actual number of EBV copies pre-capture/pre-PCR assuming a maximum 20 % loss during each of the four clean up steps. SD: standard deviation of qPCR quantity determination.

Figure 3.5 shows the ratio of viral to human load at different protocol steps for ebv8 and ebv11 with and without WGA. For ebv8, the WGA-treated sample was less enriched in EBV sequences (1.5-fold difference). After sequencing, however, a higher percentage of paired-end reads mapped to EBV from the WGA-treated sample (1.66 % versus 0.77 %, table 3.5). This is surprising and might be be due to PCR duplicates which have been removed from the mapping.

The qPCR determining human loads for the post-capture/post-PCR step failed for both ebv11 samples and ratios could not be calculated. Viral loads for these samples are 153,348 copies/ml for ebv11 without WGA and 127,589 copies/ml with WGA. This is not very meaningful in terms of enrichment success though, as concentrations differ. In the final sequencing, only 0.08 % more paired-end reads mapped to EBV in the WGA-treated sample (0.3 % versus 0.22 %, table 3.5).

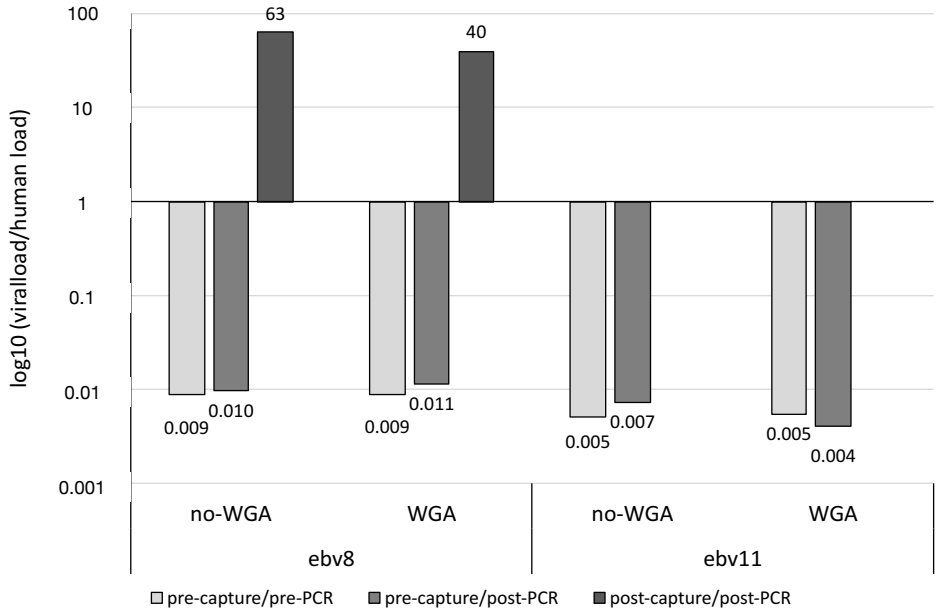


FIGURE 3.5: Ratio of viral versus human DNA at different time steps on a logarithmic scale. The two samples ebv8 and ebv11 have been used directly or treated with genome amplification (WGA) after extraction. Values at the end of the bars are the actual (non-logarithmic) mean ratios for all samples. Ratios for the post-capture step for ebv11 are missing values, not zero (see text).

Sample	Read pairs	OTR pairs	OTR %
ebv6	2,071,862	22,694	1.10
ebv7	3,099,870	30,028	0.97
ebv8	1,541,928	11,913	0.77
ebv8G	2,092,063	34,796	1.66
ebv11	2,658,858	5,785	0.22
ebv11G	2,511,284	7,586	0.30

TABLE 3.5: Sample and sequencing information for blood samples enriched without and with WGA (samples marked with a G).

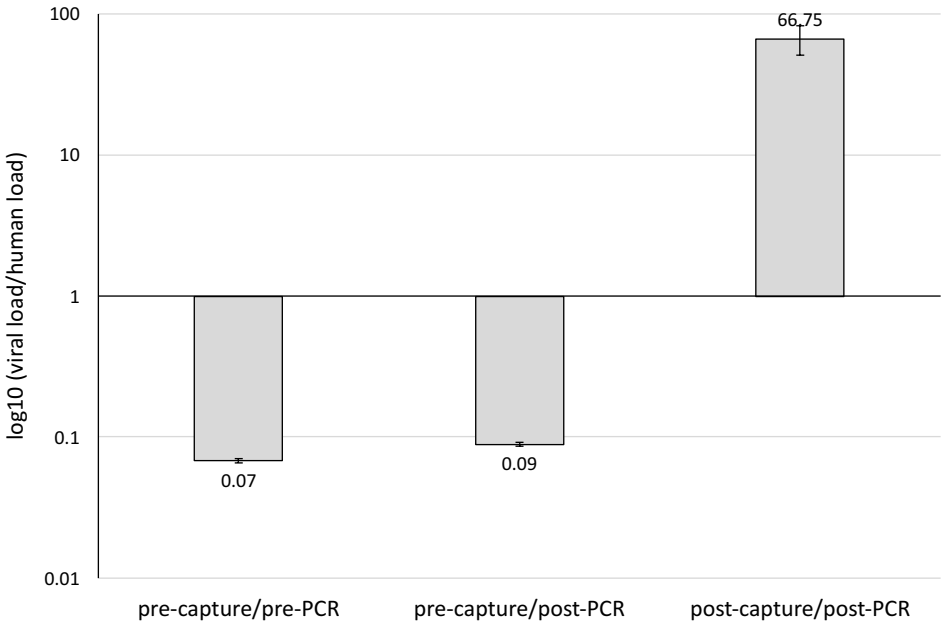


FIGURE 3.6: Average ratios of viral versus human DNA at different time steps on a logarithmic scale. Values at the end of the bars are the actual (non-logarithmic) mean ratios for all samples. Standard deviation marked by whiskers.

Figure 3.6 shows the average ratios of viral to human load of all other samples. Pre-capture, ratios were very similar across samples ($SD_{pre-PCR} = 0.0024$, $SD_{post-PCR} = 0.0028$) . The post-capture ratios had a greater variability between samples ($SD_{post-capture} = 15.72$). Ultimately, this shows that enrichment worked for all of them, but OTR output per individual sample is still fairly poor compared to other herpesviruses (table 3.5).

When comparing the OTR percentages of ebv6, -7, -8 and -11 of this SureSelect run (table 3.5) with the previous run (table 3.2), they varied. Samples ebv6, -7, and -8 had higher OTR percentages in run 2, but ebv11 had a lower value. One difference between the runs is the prior WGA of run 1, while run 2 was sequenced directly without WGA. However, another factor to account for is variability during the protocol execution. There can be variability between SureSelect runs: for example, the hybridisation step in the manual protocol, during which the samples are treated directly on a 65°C cycler, is very

sensitive to time and handling of the plates as evaporation affects its efficiency. While utmost care was taken during the preparation of the samples, the relative contribution of human error and variability likely differs between runs (i.e. practice makes perfect).

As ebv8 and ebv11 are only two samples treated in the same run with and without WGA, whether the effect of WGA solely on enrichment is positive or negative is hard to determine here. On the one hand, it is useful to amplify the amount of DNA available when dealing with scarce clinical material. On the other hand, there is a danger of introducing bias, e.g. amplification of contaminant DNA present in kits, or non-specific extension of random primers. Even though the reaction is supposed to be nonspecific, a small but significant bias in amplification for any form of WGA has been shown (Pinard et al., 2006). In a recent comparison of WGA kits and their contaminant effect, the Illustra Genomiphi V2 kit (also employed here) had the largest amount of nonspecific background reads (Thoendel et al., 2017). WGA in conjunction with SureSelect was previously tested for bias for WGS of VZV. In this study, no bias was found regarding both the consensus sequence and the population structure (Depledge et al., 2014). However, VZV has generally a far lower GC content than EBV (46 % versus 60 %), making the application of WGA amplification possibly less problematic, as it has been found that Phi29 polymerase efficiency is linked to regional GC content (Bredel et al., 2005) and leads to misrepresentation of GC regions (Arriola et al., 2007). In consequence of all of these consideration, it was decided to forego WGA if possible (i.e. if enough material available), as the benefits do not outweigh the risks.

Varying amount of RNA baits for capture

It was tested whether the carry-over of nonspecific DNA can be reduced by decreasing the amount of baits used in the hybridisation reaction. The rationale behind this is that enrichment works well in conditions where there are fewer baits than template DNA. Having more baits than template DNA, however, could lead to a large amount of nonspecific binding to non-target DNA, leading to a dilution of the OTR.

For this, DNA from three additional blood samples of immunocompromised paediatric patients was extracted and used in the 3 µg SureSelect protocol. Dependent on the total amount of DNA available after the first PCR step for each sample, different number of hybridisation reactions were set up. The standard amount of tier 1 RNA baits is 2 µl (which was estimated to contain 50k-100k copies of each individual bait, personal communication D. Depledge). Sample ebv13 was hybridised with 2 and 0.5 µl, ebv14 with 2, 1 and 0.5 µl, and ebv15 with 2 and 0.5 µl baits.

Table 3.6 lists the sample and sequencing information. For every sample, reducing the number of baits increased the OTR percentage. Using 0.5 µl instead of 2 µl increased OTR percentage from 1.3-fold (ebv14) up to 1.8-fold (ebv13). This suggests that the ratio of baits and template has an influence on efficiency and that using less baits is advantageous for WGS of pathogens from clinical samples, where target template is scarce.

Sample	Viral load [copies/ml]	Baits [µl]	Read pairs	OTR pairs	OTR %
ebv13	> 2,000,000	2	2,842,761	69,410	2.44
		0.5	3,209,553	143,600	4.47
ebv14	> 2,000,000	2	3,009,397	534,177	17.75
		1	3,172,899	656,110	20.68
		0.5	3,260,671	758,340	23.26
ebv15	1,300,000	2	3,091,208	30,771	1.00
		0.5	2,763,682	38,512	1.39

TABLE 3.6: Sequencing data of the bait dilution experiment using the 3 µg protocol.

SureSelect 200 ng protocol

In the beginning of 2014, Agilent provided a new protocol for SureSelect which allowed a lower amount of input DNA (200 ng). Principal modifications in the protocol were 1) to shear in a lower volume (50 µl instead of 130 µl), which allowed the removal of one purification step, and 2) the dilution of the adapters in order to maintain the optimal ratio for ligation.

Using DNA extracts of whole blood of immunocompetent, paediatric infectious mononucleosis patients from Japan, the 200 ng and 3 µg protocol were compared while also testing varying dilutions of RNA baits. The DNA samples were provided by Tetsushi Yoshikawa (Department of Pediatrics, Fujita Health University School of Medicine, Toyoake, Japan).

Independent of the protocol, dilution of baits improved the OTR percentage in three of four cases with an increase ranging from 1.4-fold to 3.6-fold, while the OTR percentage was retained in the other single case (table 3.7). The OTR percentage in three of four cases was increased in the 200 ng protocol (comparison of samples treated with the same bait dilutions). Only in one case was the OTR percentage lower in the 200 ng protocol than in the 3 µg protocol (P2-1246 with 1 µl baits).

Sample	Viral load [copies/µg]	Protocol	Baits [µl]	Read pairs	OTR pairs	OTR %
P2-1213	125,000	3 µg	1	1,974,835	1,157	0.06
			0.1	2,344,993	1,436	0.06
		200 ng	1	1,497,801	3,778	0.25
			0.1	2,214,579	7,548	0.34
P2-1246	14,000	3 µg	1	1,968,292	1,247	0.06
			0.1	144,649	205	0.14
		200 ng	1	2,138,754	885	0.04
			0.1	1,795,753	2,710	0.15

TABLE 3.7: Sequencing data of the bait dilution experiment using the 3 µg and 200 ng protocol.

The viral load data of these samples were provided as copies/µg of DNA, making a comparison with the OTR data dependent on viral load from previous analysed samples of immunocompromised patients hard.

While there are less viral genomes in the starting material of 200 ng DNA, the protocol has one less clean up step, saving approximately 20 % of DNA. OTR percentages being higher in most cases with the 200 ng protocol as well as when diluting baits supports the rationale that increasing the ratio of target to bait improves SureSelect performance through a more favourable hybridisation, in particular in samples with low target titres. Together, these data support the use of the 200 ng protocol and bait dilutions. It allows to use less clinical material in addition to saving reagents.

Automation system

In June 2014, the lab acquired an Agilent Bravo automation system, on which the whole library preparation including capture is performed. It also allows the preparation of up to 96 samples per run instead of the 12-16 sampling using a manual approach. 20 DNA extracts from whole blood from eleven immunocompetent, Japanese IM patients (part of the same data set as above, see table 2.1) underwent library preparation and the 200 ng SureSelect protocol with a bait dilution of 1:10 (0.2 µl) on the automation system. Final libraries were then sequenced on a MiSeq.

Table 3.8 shows the sample and sequencing data of the first runs. In a previous experiment, two samples of the same original sample set – P2-1213 and P2-1246 – were prepared manually (table 3.7. Their viral loads were approximately 125,000 and 14,000 copies/µg, respectively. Samples from this experiment here with a similar viral load are P1-812 and P4-2274 (114,000 and 139,000 copies/µg) as well as P3-2740 and P7-2315 (17,000 and 10,000 copies/µg). The OTR percentages for all four samples prepared on the automation system are several fold higher compared to the respective sample of similar viral load from the manual run (6.38 % and 16.18 % compared to 0.34 %, and 0.83 and 1.17 % compared to 0.15 %). This could be due to a more homogeneous treatment of samples, as they were prepared in parallel rather than sequentially.

These data shows that performance is at least as good as the manual protocol and support the use of the automation system for SureSelect.

Sample	Viral load [copies/µg]	Read pairs	OTR pairs	OTR %
P1-812	114,000	1,901,552	121,401	6.38
P3-2670	250,000	1,836,143	148,897	8.11
P3-2740	17,000	1,573,383	13,131	0.83
P4-2274	139,000	1,717,515	277,885	16.18
P4-2392	34,000	899,732	13,653	1.52
P5-1294	60,000	1,539,674	81,563	5.30
P5-1323	7,000	1,747,803	11,587	0.66
P6-1751	50,000	1,916,406	147,200	7.68
P6-1789	350	1,811,779	1,040	0.06
P7-2315	10,000	1,617,971	18,878	1.17
P7-2634	24,000	1,138,767	20,707	1.82
P8-414	22,000	1,814,167	282,930	15.60
P8-516	9,000	1,480,851	31,102	2.10
P9-2631	63,000	1,631,433	178,955	10.97
P9-2645	38,000	1,451,477	102,018	7.03
P10-2187	4,000	1,712,846	4,248	0.25
P10-2777	210	1,299,344	1,783	0.14
P11-871	24,000	1,703,065	242,028	14.21
P12-1026	3,000	1,428,766	25,066	1.75
P12-1078	1,000	1,743,347	1,089	0.06

TABLE 3.8: Sequencing data of infectious mononucleosis samples prepared on the automation system.

OTR vs. viral load

To demonstrate the relationship between OTR percentage (and therefore genome sequencing success) and viral input, all sequenced samples from blood extracts of immunocompromised patients (data set presented in this chapter as well as chapter 5, table 5.4 and table 5.1) were plotted (figure 3.7). The IM samples were excluded from this graph, as their viral titres were provided in a different unit.

Black symbols represent samples for which a whole genome with at least 20x coverage could be recovered, while grey symbols represent the samples for which that was not the case. Samples, for which viral loads were given as >2,000,000 are plotted at 2,000,000, but marked as a white symbol. Triangles and circles refer to the 3 µg and 200 ng protocol, respectively.

Viral load and OTR are correlated ($R^2 = 0.9$). Samples with higher viral titres result in a higher number of EBV-specific reads, independent of the protocol. According to this graph, the limit of this method lies between 20,000 and 40,000 copies/ml. For lower

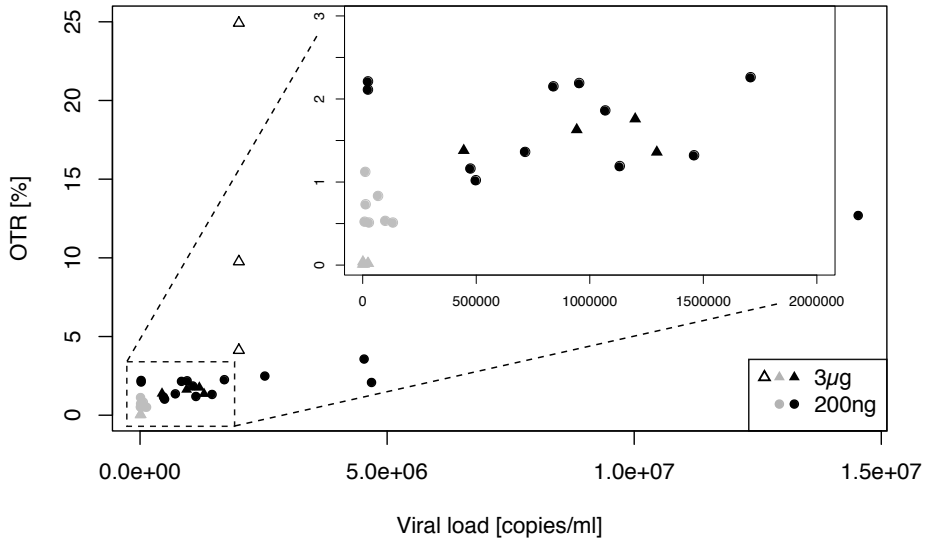


FIGURE 3.7: On target reads (OTR, percentage of EBV-specific reads) as a function of the viral load when sequencing with targeted enrichment from whole blood extracts. Grey: Samples where the whole genome could not be recovered. Black: Samples where the whole genome (>90 %) could be recovered. White: Genomes could be recovered, but no accurate viral load was given (>2,000,000); triangles: 3 µg protocol; circles: 200 ng protocol.

viral loads, enrichment and sequencing success appears stochastic and may be strongly influenced by the sample type, DNA quality and relative amount of virus present.

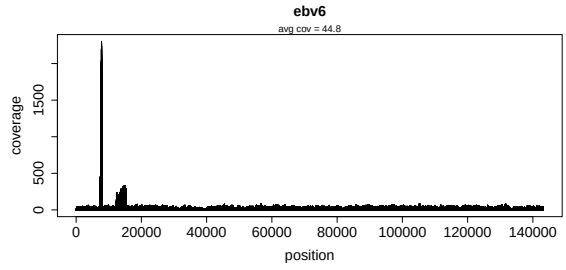
3.2.4 Final assemblies of first blood-derived EBV genomes

To obtain the final assemblies of the blood-derived EBV genomes from immunocompromised patients, reads were pooled from the different SureSelect experiments. They were processed and assembled using *de novo* assembly pipeline 1 (chapter 2). Samples ebv1-5 and -11 all had too few OTR and were discarded.

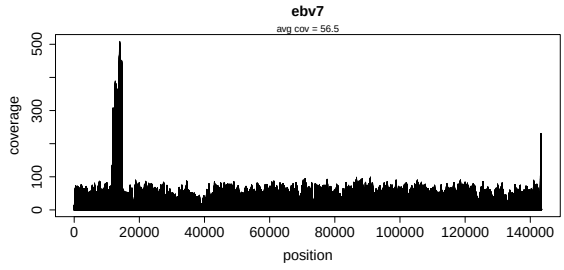
In total, seven full genomes could be recovered directly from clinical blood samples. Table 3.9 lists the sample data as well as sequencing and assembly data for the combined assemblies. As some samples have been sequenced more often than others, read numbers and depth differ between samples. The average read depth is given after removal of duplicate reads and ranged from ≈ 45 to ≈ 3700, although uneven across the genome (figure 3.8). The dominant peak observed in many of the samples is not due an over-representation of reads because of homology with human genomic areas, as a blastn search against the human nucleotide database did not deliver any significant hits; it corresponds to the viral IR1 repeat and results from reads derived from each repeat unit mapping to a smaller condensed assembled version in the primary *de novo* consensus sequence. In the final consensus sequence, this area – together with the other major repeat regions – was extended to the length of the repeat in the WT genome but masked with 'N' for subsequent analyses due to their accurate assembly being precluded by short Illumina reads.

Sample	Viral load [copies/ml]	Read pairs	OTR %	Depth	Cov
ebv6	944,000	4,150,874	1.6	45	95
ebv7	445,000	6,208,600	1.4	57	98
ebv8	1,200,000	7,277,442	1.8	129	95
ebv9	>2,000,000	5,509,174	9.8	438	100
ebv13	> 2,000,000	12,110,184	4.1	472	100
ebv14	> 2,000,000	18,894,696	24.9	3650	100
ebv15	1,295,000	11,715,202	1.4	154	99

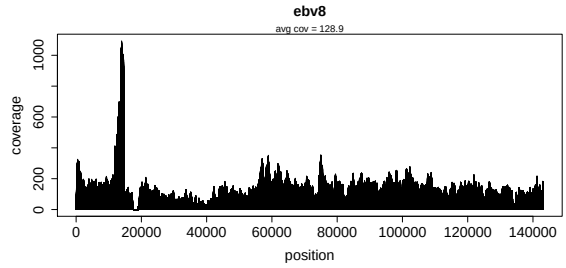
TABLE 3.9: Sample information and assembly statistics of seven blood-derived EBV genomes from immunocompromised children. Depth: Average read depth after removal of duplicated reads. Cov: Percentage of the genome that could be recovered at min. 20x (excluding repeat regions).



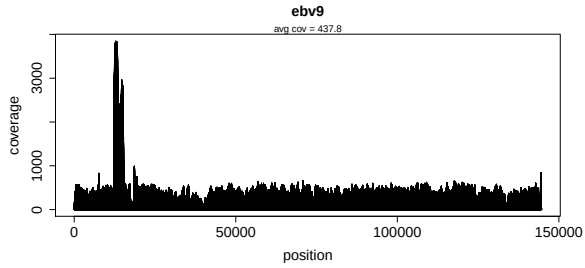
(A) ebv6



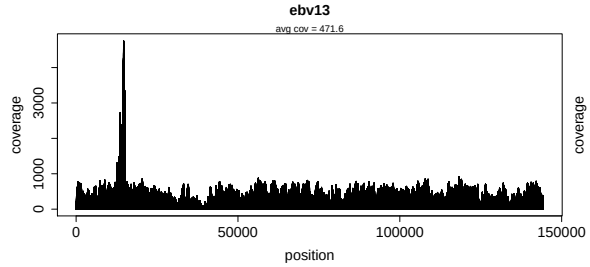
(B) ebv7



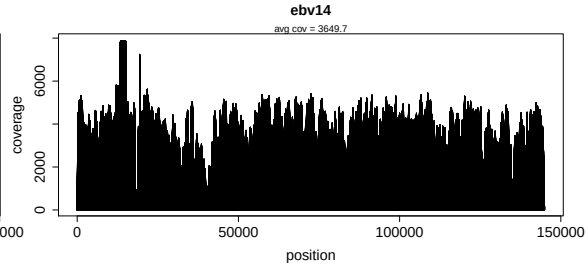
(C) ebv8



(D) ebv9

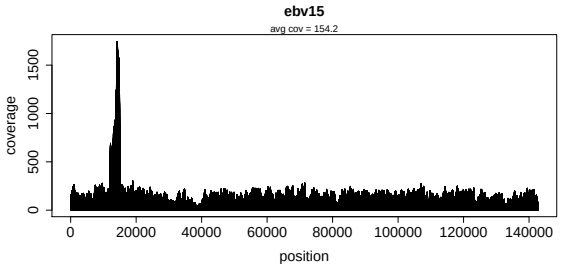


(E) ebv13



(F) ebv14

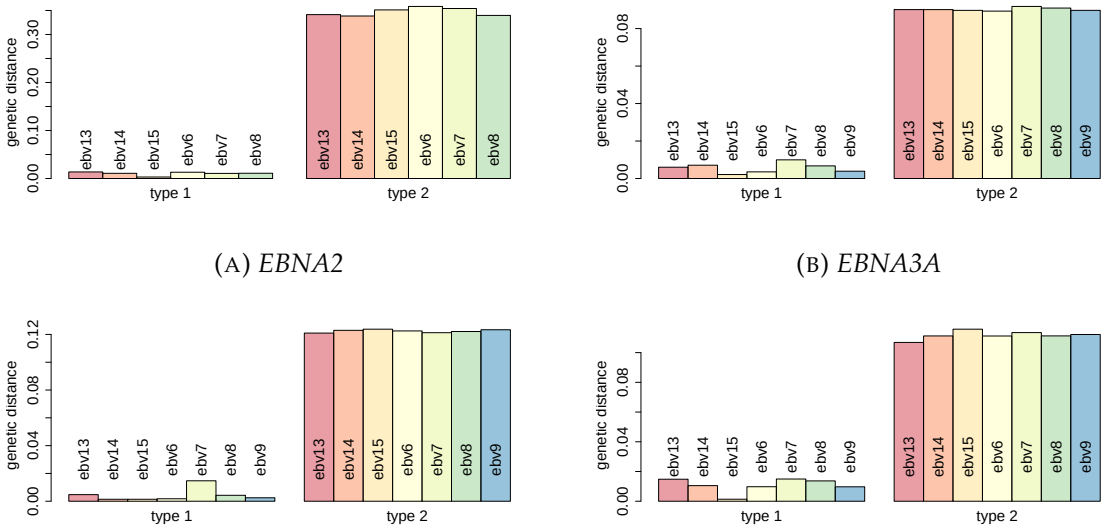
Figure continues on the next page.



(G) ebv15

FIGURE 3.8: Coverage plots of immunocompromised patient samples after duplicate removal. Mapping is done against the sample consensus sequence directly after assembly (i.e. repeat regions are not masked or considered specifically).

To discriminate as to whether samples were type 1 and type 2, the genetic distance of the typing genes *EBNA2* and *-3A-C* to the type 1 and 2 reference strains (figure 3.9) was calculated. All samples belonged to type 1 as distance for each gene was always lowest to the type 1 reference.



(A) *EBNA2*

(B) *EBNA3A*

(C) *EBNA3B*

(D) *EBNA3C*

FIGURE 3.9: Genetic distance (K80 model) of *EBNA2* and *-3* genes of samples of immunocompromised patients from the UK to type 1 (WT) and type 2 (AG876) reference sequences.

3.2.5 Comparison of blood-derived vs. tumour/LCL-derived EBV genomes

As these were the first EBV genome sequences derived directly from blood, the question arose whether they exhibit any genetic differences to LCL- or tumour-derived EBV genomes. To address this, a whole genome multiple alignment was created using these seven genomes supplemented with all published type 1 sequences (marked with a black circle in table 1.2).

Analysis of genetic variability revealed no significant differences were found in consensus sequences between the tumour/LCL-derived and blood circulating EBV genomes. They displayed a similar average number of SNPs and hotspots of diversity are in the same areas as previously described. Figure 3.10 shows the nucleotide diversity across the genome separately for blood- (n=7), tumour- (n=46) and LCL (non-malignant, n=29)-derived sequences. High diversity areas correspond to the known polymorphic genes *EBNA2*, *EBNA1*, *LMP1*. Note that the higher peaks for the blood-derived genomes is an artefact from the far lower number of sequences. This also applies for the peak around 59,000 nt, corresponding to the *BOLF1* gene, which has a large proportion of missing data in the blood sequences, reducing the number of sequences (and therefore comparisons) even further.

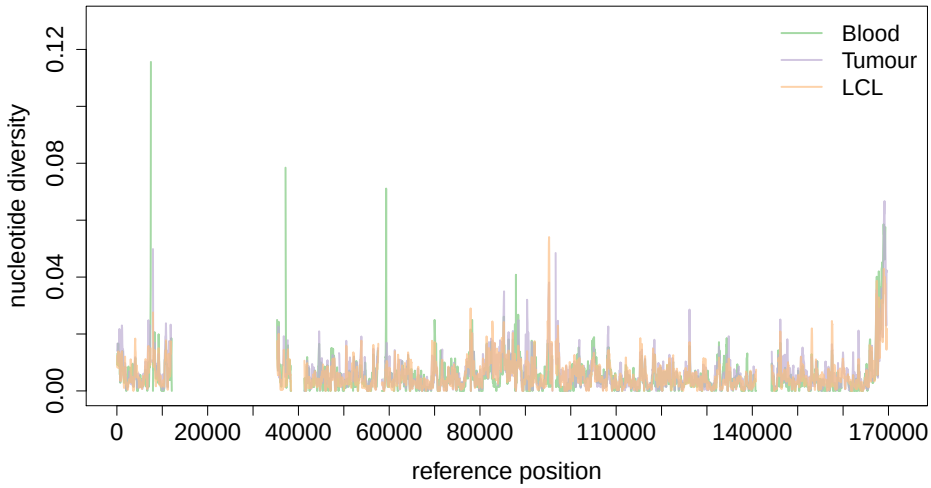


FIGURE 3.10: Nucleotide diversity calculated in sliding windows (size 200 bp, step size 50 bp) for different groups of sequences. Green: blood-derived sequences (n=7), violet: tumour-derived sequences, orange: LCL-derived sequences (from non-malignant settings). Positions refer to the WT genome. Major repeat regions are excluded.

Table 3.10 shows the average pairwise distances within group of sequences. The average pairwise distance is lowest for blood, followed by LCLs and highest in tumours. The set of tumour sequences contains a number of Chinese NPC-derived sequences which are known to differ most from other genomes. Pairwise distances have therefore also been calculated for the tumour-derived genomes for the NPC and non-NPC tumour subsets. The distance between non-NPC samples is with 0.0056 closer to both blood and LCLs, while the NPC samples are as expected for more similar to each other. This suggests that the higher average pairwise distance for all tumours results from a sampling effect (i.e. inclusion of a group of very distant sequences), and is not due to a generally higher diversity within tumour-derived versus blood-derived sequences.

No polymorphisms were found limited solely to virus circulating in blood. This is supported by the finding that sequences derived from blood did not cluster separately in

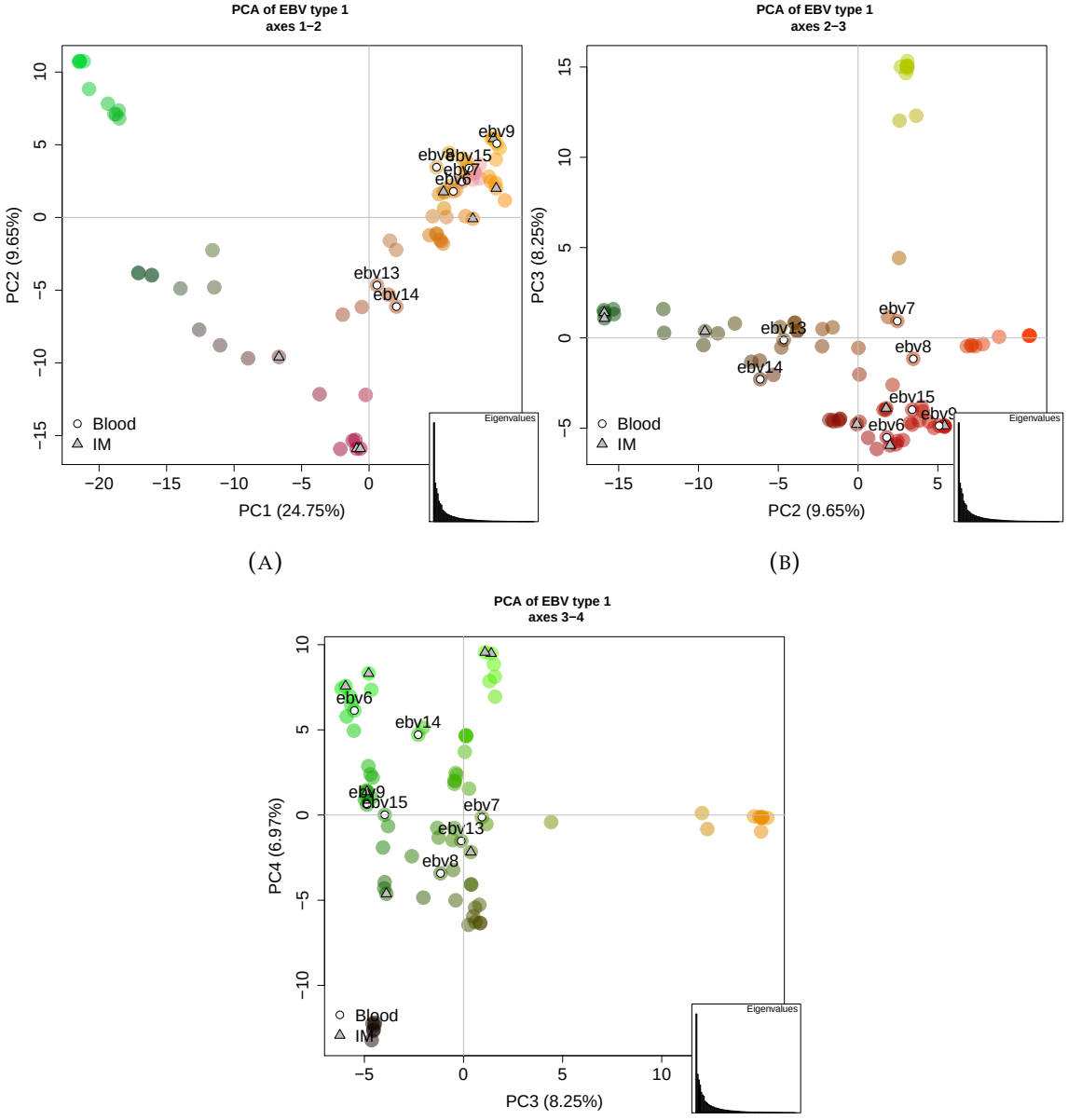
	Dist	SD
Blood	0.0051	0.0014
Tumour	0.0063	0.0023
LCL	0.0055	0.0024
Tumour - NPC	0.0033	0.0028
Tumour - non-NPC	0.0056	0.0021

TABLE 3.10: Average pairwise distances between genomes derived from blood, tumours and LCLs (non-malignant settings), calculated using the K80 distance model. SD: standard deviation.

a whole-genome based PCA (figure 3.11, blood-derived samples are marked by a white circles). Here, principal components (PC) 1 and PC2 discriminate between Asian and non-Asian genomes, whereas PC2 and PC3 as well as PC3 and PC4 discriminate between some of the African, and the Western and Asian genomes. (The same PCA is shown in supplementary figure A.1 with all sampled labelled by their geographic origin.) Also in higher principal components, accounting for less percent of total variation, samples never clustered based on their compartmental origin.

The latency genes *LMP1*, *LMP2*, *EBNA1*, *EBNA2*, and *EBNA3A-C* are functionally important in establishing and maintaining latency. In addition several are oncogenes. The PCA of these gene sequences revealed no discernible difference between blood-derived and tumour-derived EBV genomes. Figure 3.12 (left panel) shows that the sequences recovered directly from whole blood are never specifically discriminated from other genomes. Instead, they frequently cluster with other LCL- or tumour-derived genomes. The right panels of figure 3.12 show unrooted Neighbor joining (NJ) trees of these genes. They do not depict the evolutionary relationship between the samples, but should be interpreted as a simple distance-based clustering to complement the PCA plot. The figure depicts only the first two PCs, but the non-segregation was also observed for higher components examined.

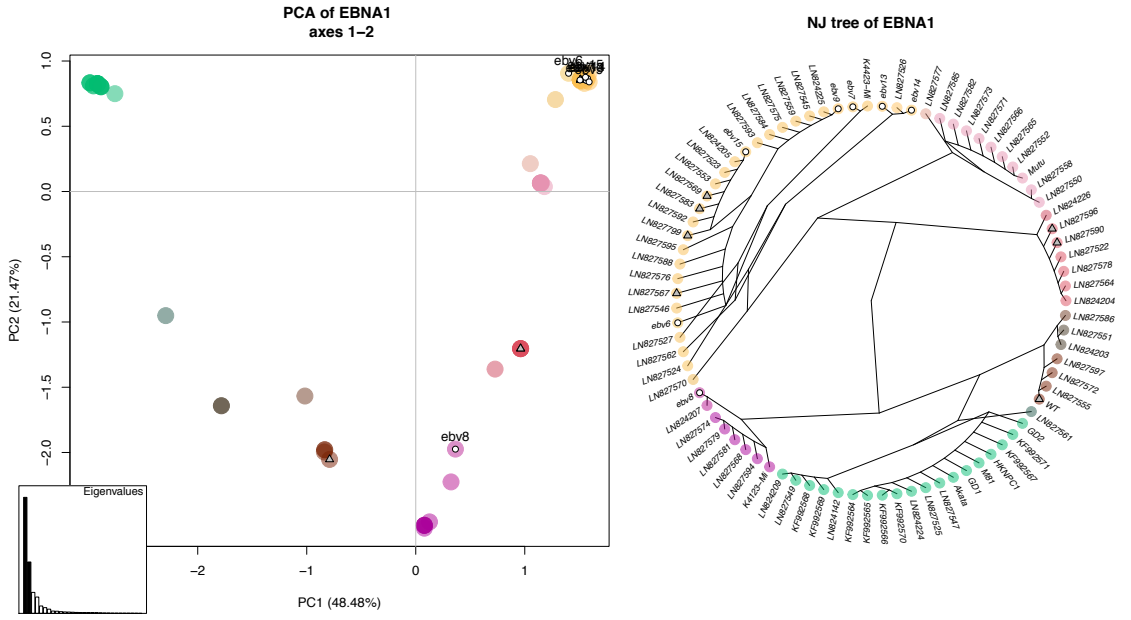
In figures 3.11 and 3.12, samples derived from IM patients (usually transformed LCLs) are also specifically marked (grey triangles) as examples of isolates from non-cancerous pathologies. However, LCL-derived IM samples could never be discriminated (together with the blood-derived samples or on their own) by the PCA.



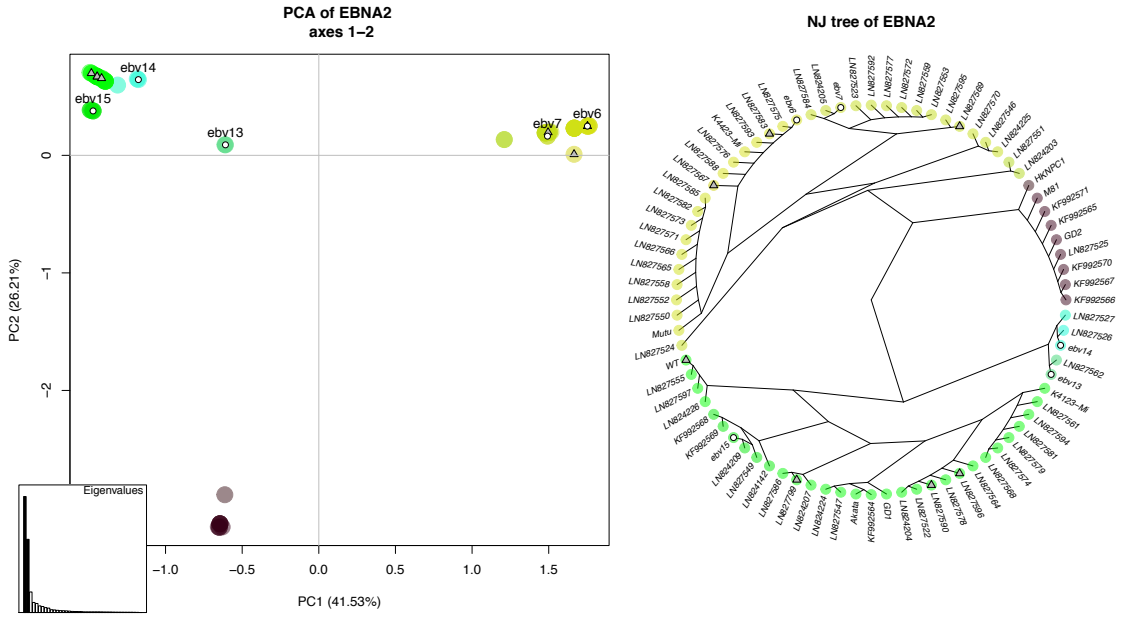
(A) (B)

(C)

FIGURE 3.11: PCA of whole genome sequences for type 1. Specifically marked are blood-derived samples (white circles) and isolates from IM patients (grey triangles). The bar plot in every corner shows the eigenvalues for every principal component (PC), i.e. their influence on the variation of the data. The colouring is an RGB translation of each sequence's PCs scores. 3.11a shows the PC1 against PC2, 3.11b shows PC2 against PC3 and 3.11c shows PC3 against PC4. The histogram in the corner of each PCA plot shows the distribution of eigenvalues (i.e. contribution) of the PCs from left to right, with plotted PCs being highlighted.

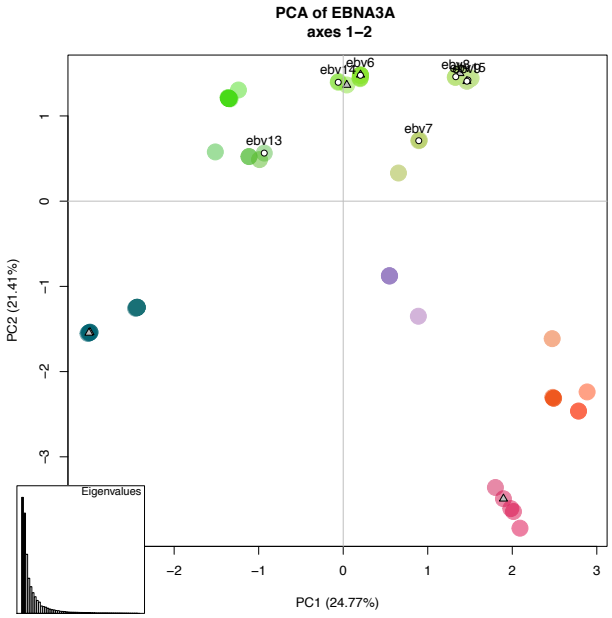


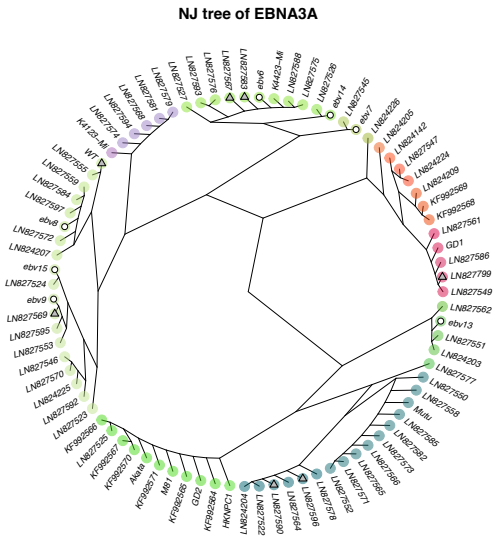
(A) EBNA1



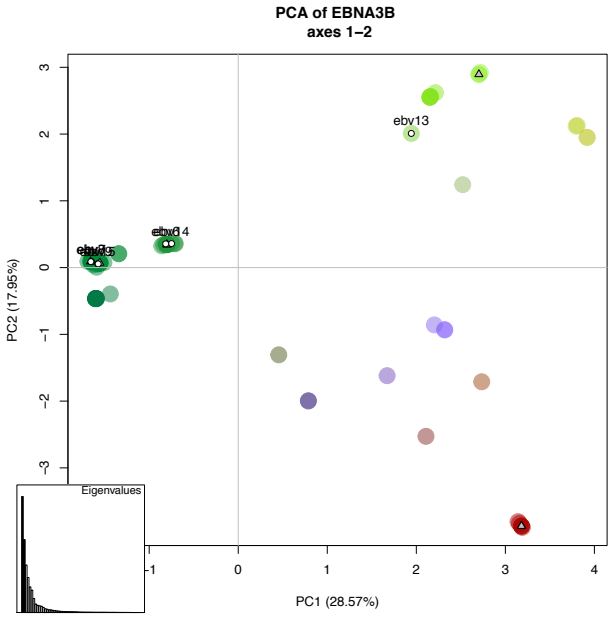
(B) EBNA2

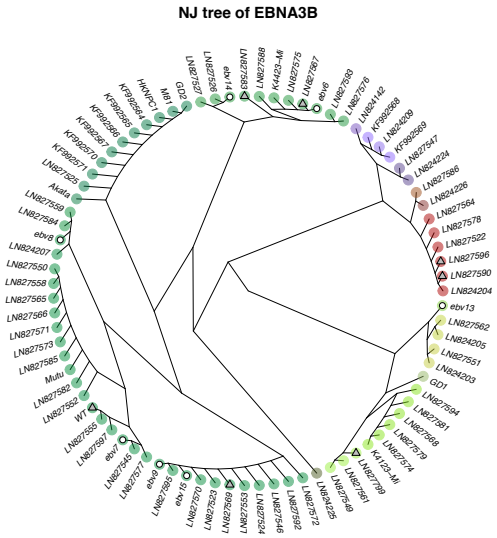
Figure continues on the next page.





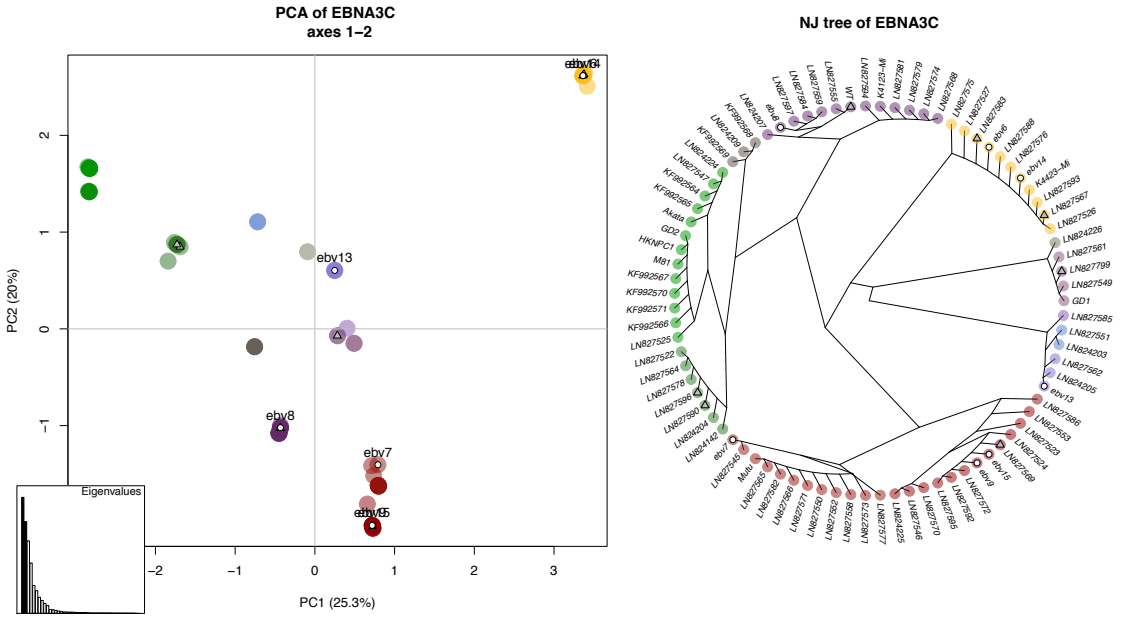
(C) EBNA3A



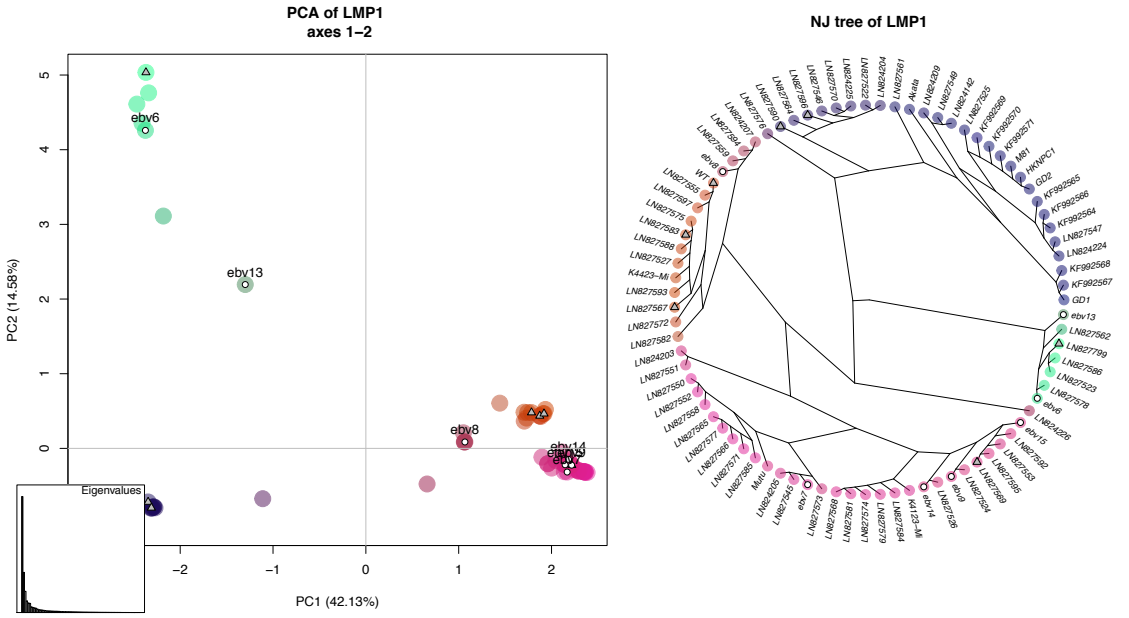


(D) EBNA3B

Figure continues on the next page.

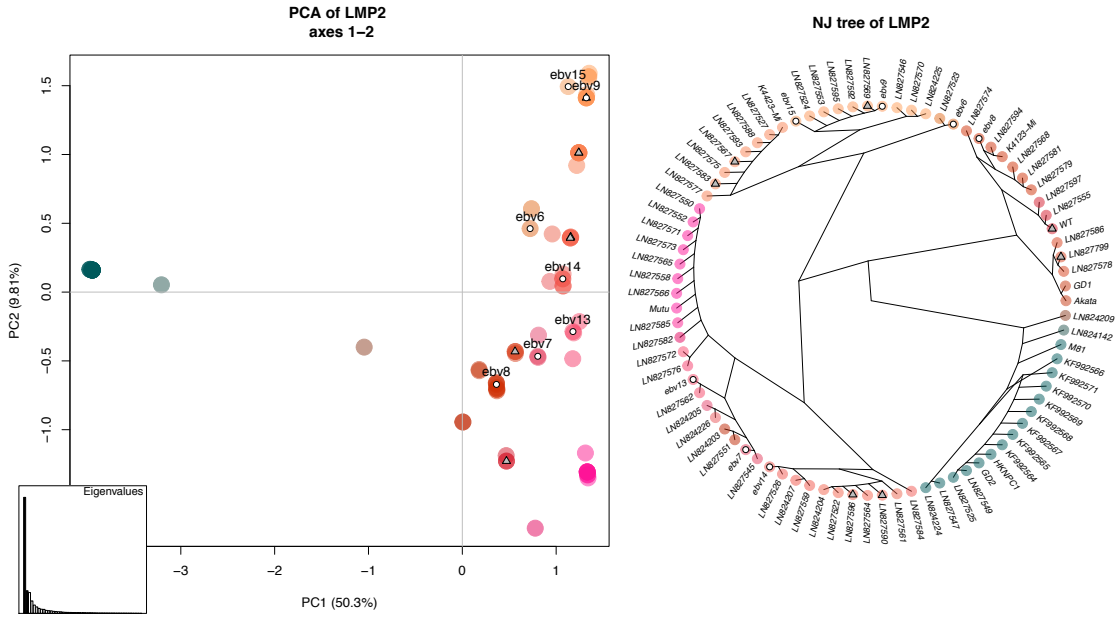


(E) EBNA3C



(F) LMP1

Figure continues on the next page.



(G) LMP2

FIGURE 3.12: Left Panel for subfigures 3.12a-3.12g: PCA results for the first two principal components (PCs) of the latency genes. The first two PCs do not distinguish between EBV sequences from natural infection vs. tumours, nor do any of the higher principal components (data not shown). Specifically marked are the non-cancerous samples, i.e. blood-derived samples (white circles) and isolates from IM patients (grey triangles). The bar plot in every corner shows the eigenvalues for every principal component (PC), i.e. their influence on the variation of the data. The colouring is an RGB translation of each sequence's PCs scores. The histogram in the corner of each PCA plot shows the distribution of eigenvalues (i.e. contribution) of the PCs from left to right, with plotted PCs being highlighted.
Right Panel for subfigures 3.12a-3.12g: Simple clustering based on a Neighbor joining (NJ) tree of the latency genes.

3.3 Discussion

3.3.1 Optimisation of SureSelect for EBV sequencing from whole blood

Here, the SureSelect protocol was optimised to enrich DNA extracts from blood for EBV prior to WGS. The method was first described for viral genome sequencing by Depledge et al., 2011. The sequences described here are the first EBV whole genome sequences sequenced directly from blood. Other published genomes have been sequenced either from tumour tissue or cell cutures - with the exception of one saliva sample (Palser et al., 2015) - where viral titres are generally a lot higher.

The SureSelect protocol is a nascent method and has been repeatedly updated throughout the study. The data presented here indicate that protocols using 200 ng of input DNA produce better outcomes than those using 3 µg, despite having less total viral genomes in the starting material. This is beneficial as it saves precious clinical material. More

importantly, it was found that diluting the RNA baits during the hybridisation-based enrichment increases the proportion of EBV-specific reads, particularly when viral loads are low. Both factors should result in a increasingly favourable ratio between baits and target DNA during hybridisation, which allows for more efficient binding and reduces the carry-over of non-target sequences. This has an additional advantage of reducing costs associated with library preparation.

Independent of the protocol used, the success of target enrichment is highly dependent on the viral load (figure 3.7), with a higher proportion of viral DNA leading to more EBV-specific reads in the final library. The current minimal threshold of detection found in this work lies between 20k-40k EBV genomes/ml. This threshold is less for other herpesviruses (personal communication with J. Breuer), possibly due to intrinsic features of the viruses' genomes (e.g. GC content). In the future, this might be further improved to a point through improved reagents. The threshold of detection is also partly linked to the technical limits within the protocol (i.e. loss during purification and statistical distribution of correctly ligated adapters). However, this also highlights the importance of knowing exactly how much virus is present in the starting sample, in particular if receiving samples from other institutions, in order to choose the appropriate protocol and sequencing platform.

Generally, coverage was found to be uneven across the genome. This is partly due to the intrinsic nucleotide composition of the EBV genome. For example, the overall GC-content is around 60 % (Kwok et al., 2012; Lei et al., 2013). But in some regions, GC content exceeds 80 %, which can reduce efficiency and bias negatively the PCR steps involved, and make it more likely that independent of the sample quality, certain genomic regions might be missed.

An additional source of variability in outcome could be the genetic distance of the strains present in the sample from the references used in bait design. For this reason, the EBV RNA bait set was updated in 2015, the new design taking into account the increased variation observed in more recent genome sequencing studies (see chapter 2).

Independent of these intrinsic factors, OTR outcome can vary between runs and between samples due to human error and sequential processing, especially during critical steps. In this regard, the automation system reduces the variation during the preparation allowing for greater consistency in results. Moreover, it has the advantage of allowing high throughput preparation of samples, saving both cost and time.

3.3.2 Differences between EBV genomes derived from blood versus tumours and LCLs

Previously published EBV genomes have, with a single exception, been derived from either tumours or LCLs (Palser et al., 2015). However, establishing an LCL may introduce a bias towards viral strains that most effectively immortalise B cells. Additionally, passaging virus may introduce new mutations. Moreover, tumour-derived genomes may contain genome alterations that are not representative of naturally circulating strains, either due to selection processes for strains that are (more) oncogenic, or because the virus

persists now in an altered cell environment where selection pressures differ (e.g. replication via massive clonal expansion of B cells rather than lytic replication).

In the analyses presented here, however, no blood-specific genetic variations were detected and the PCA analysis did not identify any discernible differences between EBV genomes derived from blood versus tumours/LCLs (on the SNP level). This indicates a lack of detectable differential selective pressure in either cell culture, tumour or naturally circulating virus. Moreover, widening the subset to genomes most representative of natural infection (i.e. including IM-derived LCLs) did not change the observed results either. Instead, these results confirm the previous finding of geographical segregation through the three first principal components (Palser et al., 2015). The geographic signals, however, might also override signals relating to compartmentalisation, and even sampling of different regions and compartments would be necessary to answer these questions. In addition, variation between compartments due to different modes of replication, might be more likely reflected on the minority level. This is further explored in chapter 5.

The same rationale holds for the identification of malignancy-associated variations. As many strongly EBV-associated malignancies show geography-dependent incidence rates (e.g. BL, NPC), it is hard differentiate between variation due to geography versus malignancy. However, this work has established EBV sequencing directly from blood, enabling whole genome studies to compare both tumour and blood-derived genomes from specific high incidence areas, allowing the disentanglement of these variables and the identification of potential disease associated variations.

Chapter 4

Comparative genomic analysis of world-wide EBV strains

4.1 Introduction

EBV is a world-wide distributed virus infecting only humans, whose divergence time from other γ-herpesviruses has been estimated to be 90-100 million years ago (McGeoch et al., 1995). This indicates a close coevolution for a long period of time with humans and their related primates and ancestors.

Whole genome sequencing of viruses gives comprehensive information about diversity and consequently can also be used to determine the phylogenetic relationships between samples, identify areas of the genome under selection and elucidate how the virus population is structured.

The first EBV genome, strain B95-8, was published in 1984 (Baer et al., 1984). It was isolated from a North American infectious mononucleosis patient, and served as the backbone of what is used as the reference genome for EBV now, by filling a non-representative 11 kb deletion with a segment derived from the Raji sequence in 2003 (de Jesus, 2003). This marks the beginning of actual EBV genome sequencing, as more and more sequences have slowly been published since then. GD1 and GD2, both from South Chinese NPC patients, were published in 2005 and 2011, respectively, with GD2 being the first strain to have been sequenced by NGS technology. AG876 was the first type 2 genome, derived from a Ghanian Burkitt's lymphoma in 2006 (Dolan et al., 2006). In 2012, two Burkitt's lymphoma derived genomes were sequenced, Akata from Japan and Mutu from Kenya (Lin et al., 2012). In 2013, four more genomes were published: two sLCL derived sequences from the USA, K4123-Mi and K4413-Mi, and a NPC isolate from Hong Kong, M81 (Lei et al., 2013; Tsai et al., 2013) and C666-1, a NPC cell line (Tso et al., 2013).

Finally, the introduction of target capture (Depledge et al., 2011) allowed a dramatic increase of genomes, with nine more NPC derived genomes published in 2014 (Kwok et al., 2014) and 71 genomes from various geographic origins and malignancies in 2015 (Palser et al., 2015). Moreover, three nearly complete genomes were constructed using the data of the 1000 Genomes Project (Santpere et al., 2014) and 12 Burkitt's lymphoma isolates from Ghana, Brazil and Argentina (Lei et al., 2015). The large number of genome

sequences now available from various geographic origins as well as different tumour and tissue types allows a comprehensive comparative analysis.

In this chapter, results of a comparative genomics analysis of seven whole genome EBV sequences derived from blood described in the previous chapter 3 as well as 76 type 1 EBV genome sequences which became available in Genbank at the time of this study are presented.

There are several questions: How are genomes altered by evolutionary forces? Are there biological correlates of these evolutionary forces? What is the phylogenetic relationship between strains and how is this related to population structure?

4.2 Results

4.2.1 Data set

The data set consisted of 76 type 1 EBV whole genome sequences available in Genbank (marked sequences in table 1.2), comprising samples from various geographical regions and pathologies. It was further increased by sequencing EBV from whole blood extracts of pediatric, immunocompromised patients from the UK. In total, seven full EBV genomes were recovered directly from blood (see chapter 3, table 3.9).

All analyses presented in this chapter have been conducted on consensus genome sequences. The major repeat regions FR, IR1, IR2/NotI, IR3, IR4/PstI and TR have been masked throughout the analysis.

In total, the data set is comprised of sequences derived from circulating and oncogenic virus genomes from Europe, North America, Australia, Africa and East Asia.

4.2.2 Genome-wide recombination analysis

Ascertaining the evolutionary relationship between the seven sample sequences and previously published sequences by calculating a phylogenetic tree is difficult. A PHI-test (Bruen, Philippe, and Bryant, 2006) found significant evidence for recombination ($p < 0.05$) and a genome-wide PHI-profile scan revealed areas of significant recombination throughout the genome (figure 4.1). The presence of numerous parallel edges between branches in a split network confirms this, as they depict incompatible or ambiguous signals within the data (figure 4.2). Consequently, recombination has to be carefully considered and whole genome tree topologies are likely to be affected by it.

A useful way to assess recombination on a greater scale is to consider linkage disequilibrium (LD), i.e. the correlation between the occurrence of polymorphisms at different loci in the genome (Haydon, Bastos, and Awadalla, 2004). Two loci are considered to be in LD when they occur together more often then would be expected from a random distribution of allele frequencies. There are several factors influencing LD, e.g. genetic linkage (due to physical proximity), selection and recombination. LD will be lower for example if there is recombination occurring between two sites. As it is more likely for recombination to occur between distant sites, there is a negative relationship between estimated LD of two biallelic sites and the physical distance between them (Balding, Bishop, and Cannings, 2007).

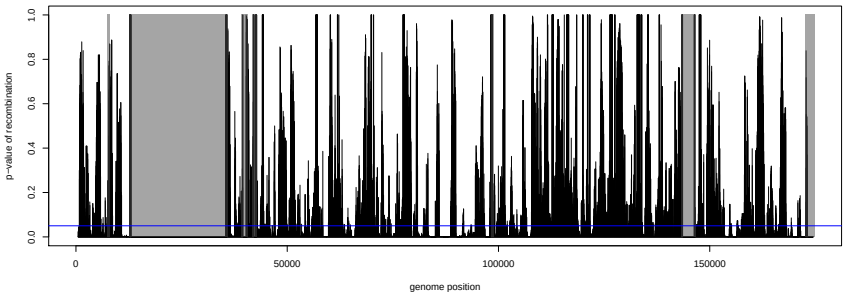


FIGURE 4.1: Profile plot of p-values of the PHI-test across the genome. The PHI-test was conducted in windows of 1000 bp with a step size 25 bp for all type 1 sequences. The blue line marks the significance threshold of $\alpha = 0.05$. In other words, all peaks lower than the blue line mark areas of the genome that show evidence of recombination. Repeat regions are marked in grey.

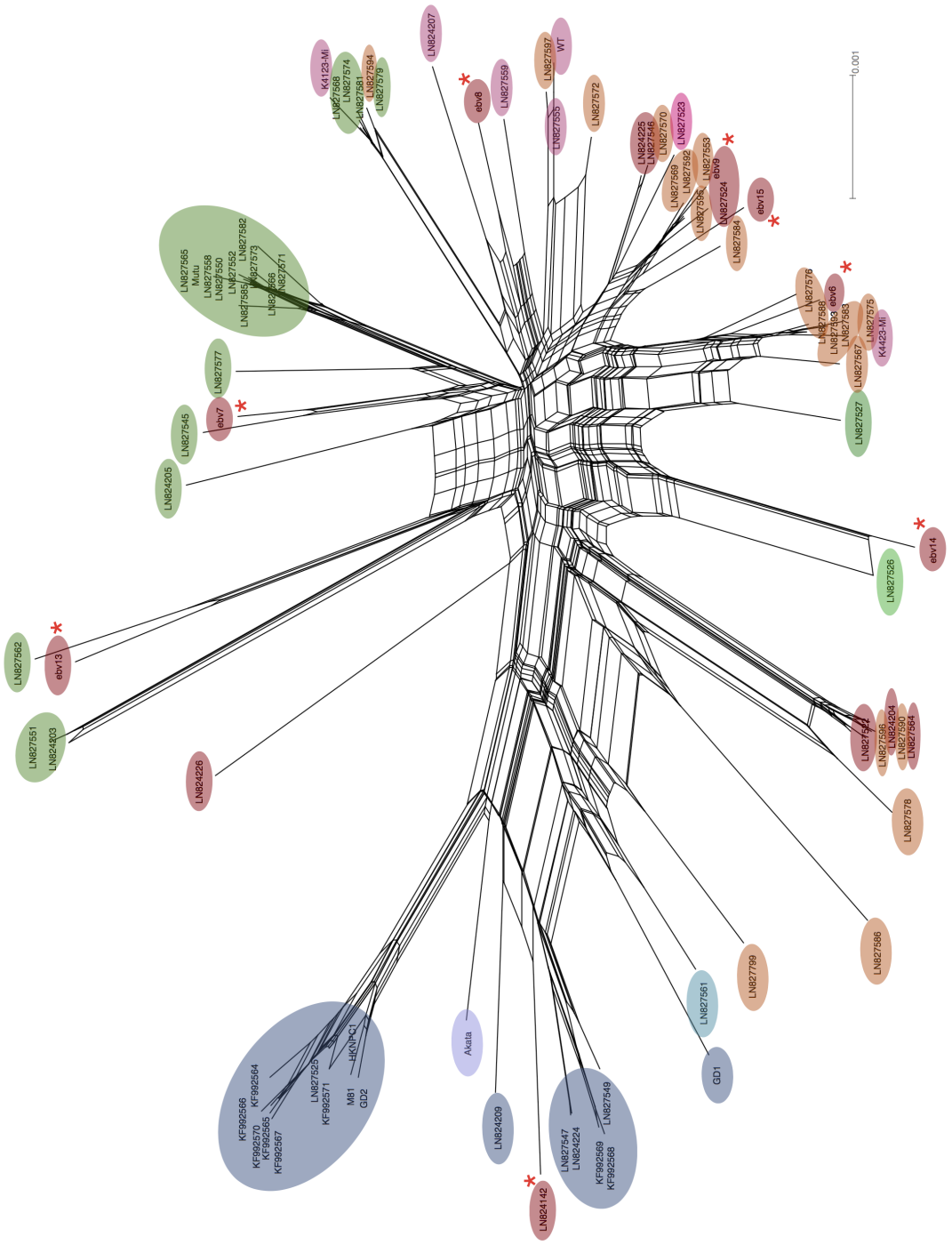


FIGURE 4.2: Split network of the whole genome alignment of type 1 sequences. Taxa are colour-coded based on their geographic origin. Blue: Asian sequences; Green: African sequences; Red and Pink: Europe and America; Orange: Australia. Genomes derived from blood and saliva are marked with a red *.

Evidence of genome-wide linkage-disequilibrium

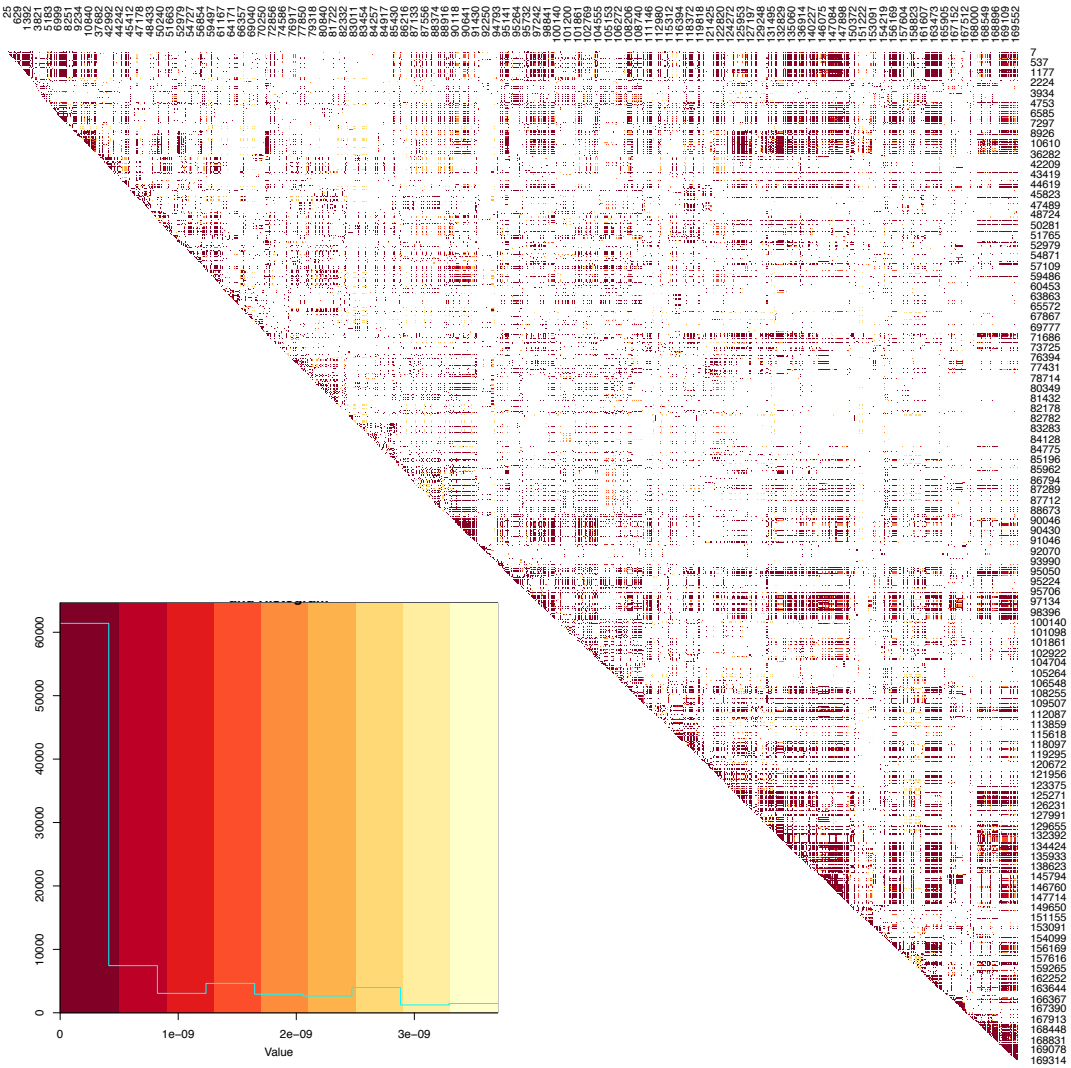


FIGURE 4.3: Heatmap displaying the significant Bonferroni-corrected p-values of Fisher's Exact test as measure of LD between biallelic sites that are at least once in LD with another site. Insignificant results ($p < 0.05$) is white. The darker colour indicates smaller p-values. Axis values refer to coordinates of every 100th biallelic site.

Figure 4.3 shows the level of LD throughout the genome as measured by Fisher's Exact test for all biallelic sites. In total, 88,787 pairs of loci were found to be in LD, which are comprised of 1,857 individual sites of the 5,190 biallelic sites analysed. There are 81,525 pairs with at least one site being located in an ORF and 49,718 pairs where both sites fall within ORFs. Of these, there are 8,038 pairs where both sites are synonymous, 23,172 pairs with at least one nonsynonymous site and 18,508 pairs where both sites lead to nonsynonymous changes.

Figure 4.4 depicts the distribution of (corrected) p-values as a measure of LD depending on the distance between two linked sites. According to this, our test measure is not

overly dependent on distance between two linked sites, i.e. we find equally statistically "strong" LD between sites very far apart as well as closer together. Only very proximate sites (in 0-1 kb distance of each other) show a noticeably lower distribution of p-values.

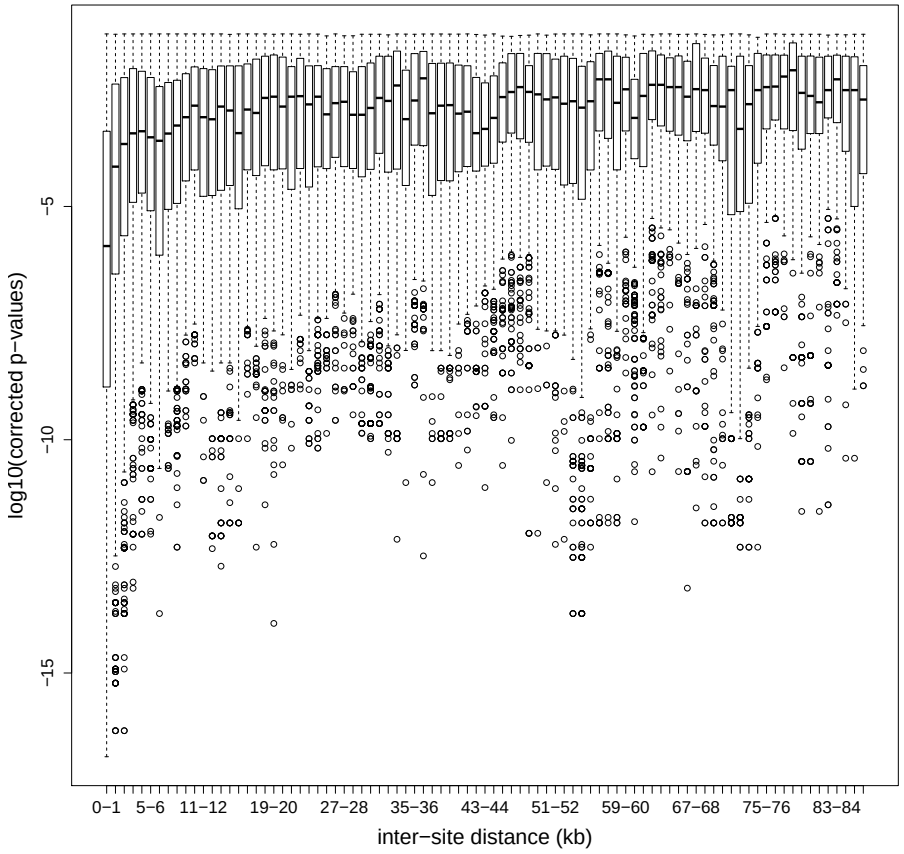


FIGURE 4.4: Distribution of p-values of Fisher's Exact test for all pairs in LD over site pair distance. The distance between two sites in LD has been binned into distance classes of 1 kb size.

From figure 4.3 it can be seen that LD is being detected throughout the genome, even between sites very far apart. This long-range LD is counterintuitive given the evidence of pervasive recombination (figure 4.1). Focusing on the subset of sites that are in LD with each other (n=1,857), recombination networks (supplemental figure B.1, B.2, B.3) and PHI-test (all sites in LD: $p < 0.05$, all nonsynonymous sites in LD: $p < 0.05$, all synonymous sites in LD: $p < 0.05$) still gave evidence for recombination occurring within these subsets. Similarly, filtering the alignment for pairs of SNPs that are strongly in LD with each other based on a p-value threshold, did not diminish the signal of recombination (PHI-test for sites in LD with thresholds smaller than 0.01, 0.02 and 0.00005: $p < 0.05$; supplemental figures B.4, B.5, B.6).

Population structure of EBV

I therefore investigated whether this pattern of linkage can be explained by an underlying population structure, potentially of mixed genetic background (i.e. allowing for recombination), using the software *structure* (Pritchard, Stephens, and Donnelly, 2000).

This software tries to cluster loci into a predefined k number of clusters (populations), while allowing for individuals to have an admixed genetic background. The best fitting number of populations has been determined using Evanno's method (Evanno, Regnaut, and Goudet, 2005). According to this *ad hoc* method, the number of clusters is chosen such, that the statistic Δk is maximal. This is depicted in supplemental figure B.8, which shows the 2 clusters best describe the data. Figure 4.5 shows the results for different subsets of biallelic sites assuming a population number of $k = 2$.

Using all biallelic sites throughout the genome (panel A in figure 4.5), all isolates from Asia belong to one population (blue), while the majority of African and Western isolates

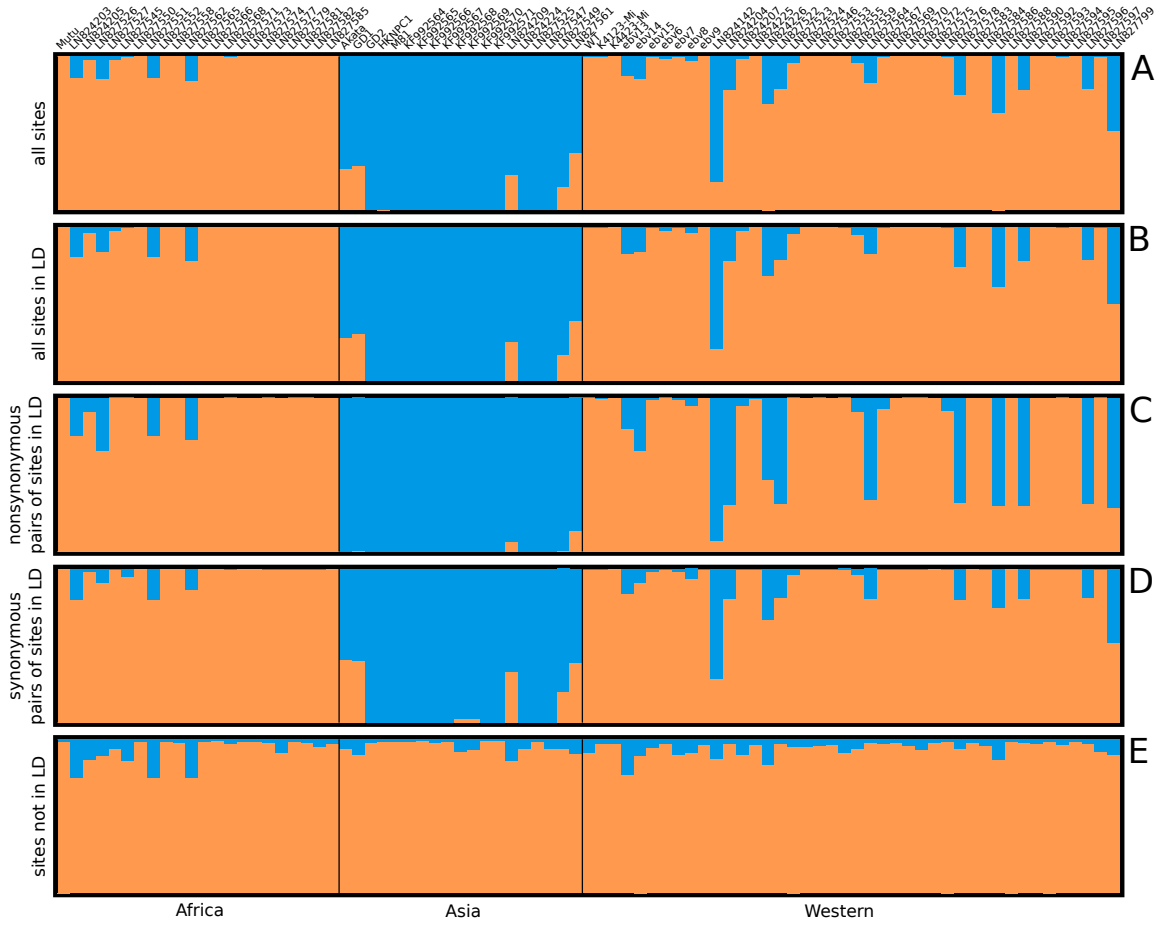


FIGURE 4.5: Population assignment for all genome sequences assuming a population number of $k = 2$ for different subsets of sites. Every bar represents a strain that has been preassigned to either "Africa", "Asia" or "Western" (comprised of American, European and Australian isolates). The colouring of the bars represents the proportion of the input sites that have been assigned to a certain population.
A) all sites; B) all sites in LD; C) nonsynonymous pairs of sites in LD; D) synonymous pairs of sites in LD; E) sites not in LD.

belong the other one (orange), suggesting the existence of an Asian virus population and a population of viruses that is spread throughout the rest of the world. A few isolates, however, have been partly assigned to the Asian population. This notably includes LN824142, a saliva-derived isolate from a healthy individual in the UK, and LN827799, an IM-isolate from Australia.

Restricting the data set to all biallelic sites in LD (panel B) does not change the proportional assignment of isolates to populations, and neither does it in the subset of pairs of synonymous sites in LD (panel D). But pairs of sites in LD that result in nonsynonymous changes (panel C) give a slightly different picture: The majority of nonsynonymous SNPs in LD found in the two previously mentioned sequences (LN824142, LN827799) as well as a few others, all of which showed an admixed genetic background, has been assigned to the Asian population. These other isolates are four Hodgkin's lymphoma isolates from the UK (LN824204, LN824226, LN827522, LN827564), two PTLD samples (LN827578, LN827586) and three further IM samples (LN827590, LN827596, LN827799) from Australia.

On the other hand, polymorphic sites not in LD with any other site do not show evidence for a defined population structure (panel E).

In order to see whether there might be a finer structure within the non-Asian population, that is being masked by the stronger signal of the Asian isolates, structure analysis has been repeated with Asian sequences removed. However, it was not possible to determine a suitable number of clusters based on the k range tested, indicating that the data does not contain structure or is too subtle or disrupted to be detected.

Analysis of linked genes

It is interesting that the majority of nonsynonymous polymorphisms in LD occurring in the recombinant "Western" isolates are assigned to the blue, supposedly Asian cluster, as selection can act primarily on polymorphisms that result in amino acid changes. I therefore wanted to explore this subset of sites further.

There are no obvious regions of the genome that are excluded in terms of the occurrence of linked sites (in general and nonsynonymous, see supplemental figure B.7). Figure 4.6 shows the top 1 % of linked ORFs, meaning those which contain the most nonsynonymous SNPs in LD with each other. Many of these genes are known to be antigens (marked by an asterisk). This led to the hypothesis that adaption to the host immune system and maintenance of a certain subset of variation might play a role in the genetic structure observed.

To test this, the data set of genes was divided into two sets: genes that are known to code for immunogenic (IG) proteins and those that do not (NIG). This list is based on the Immune Epitope Database (IEDB) (*Immune Epitope Database*) with restriction to those antigens whose epitopes have been confirmed by at least two studies and are listed in table 4.1. Nonsynonymous sites within these ORFs belonging to IG are more often in LD with each other than would be expected if a uniform distribution of links across all genes is assumed ($p < 2.2e{-}16$, Chi-square test, figure 4.7a), even when excluding links

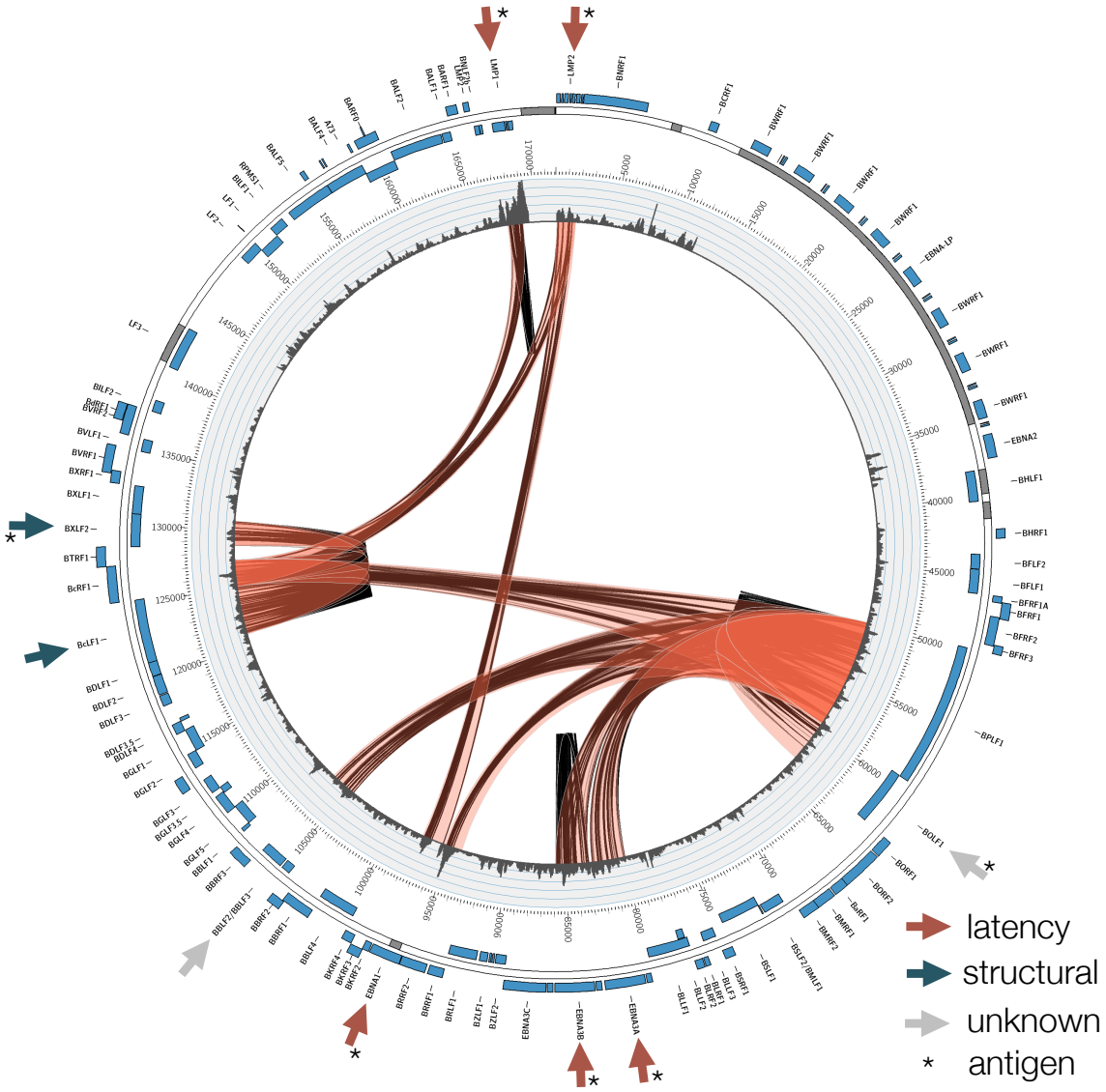


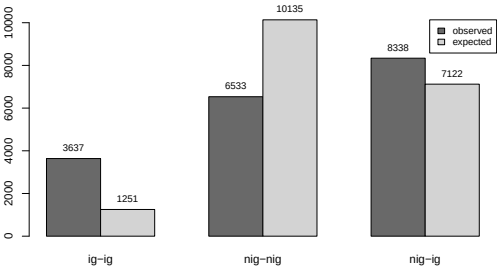
FIGURE 4.6: Genome map depicting the EBV ORFs in blue on the outside. The grey bars mark the repeat regions, which have been excluded from analysis. The outer track depicts the nucleotide diversity. The connection across the genome mark the top 1 % of ORFs that are most often linked via nonsynonymous sites. The red ribbons outline the ORF, while the indivual black lines mark the specific pairs of sites in LD. The colour-coded arrows relate to the encoded protein's function. The asterisk marks those ORFs whose products contain epitopes.

between proximal SNPs (figure 4.7b). Similarly, genes belonging to NIG are less often linked with each other.

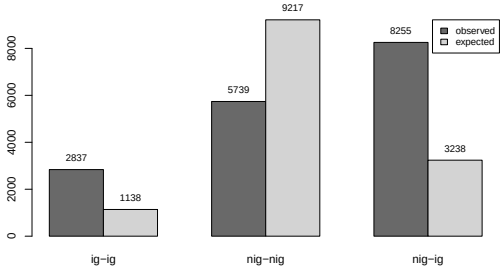
As sites are linked with each other across the whole genome. i.e. SNPs (and ORFs) are not only linked to one but several other SNPs (and ORFs), I sought to study this interconnectedness with a graph theoretical approach (see Methods). The resulting gene network consisted of 73 genes, 19 of them belonging to IG and 54 belonging to NIG, respectively. Edges were weighted based on a linkage score (see Methods). This linkage score is significantly higher for edges between genes both belonging to IG, than between

Protein	ORF	# of epitopes	# of references
Envelope glycoprotein B	BALF4	1	2
DNA polymerase catalytic subunit	BALF5	1	6
Putative BARF0 protein	BARF0	1	2
Capsid protein VP26	BFRF3	1	3
Envelope glycoprotein GP350	BLLF1	1	3
mRNA export factor ICP27 homolog	BMLF1	3	103
DNA polymerase processivity factor BMRF1	BMRF1	5	16
Major tegument protein	BNRF1	2	4
Protein BOLF1	BOLF1	1	2
Replication and transcription activator	BRLF1	7	44
Envelope glycoprotein H	BXLF2	3	10
Trans-activator protein BZLF1	BZLF1	15	96
Epstein-Barr nuclear antigen 1	EBNA1	52	163
Epstein-Barr nuclear antigen 2	EBNA2	5	16
Epstein-Barr nuclear antigen 3	EBNA3A	18	209
Epstein-Barr nuclear antigen 4	EBNA3B	15	108
Epstein-Barr nuclear antigen 6	EBNA3C	21	103
Latent membrane protein 1	LMP1	9	55
Latent membrane protein 2	LMP2	26	216

TABLE 4.1: List of 19 genes considered to code for immunogenic proteins.



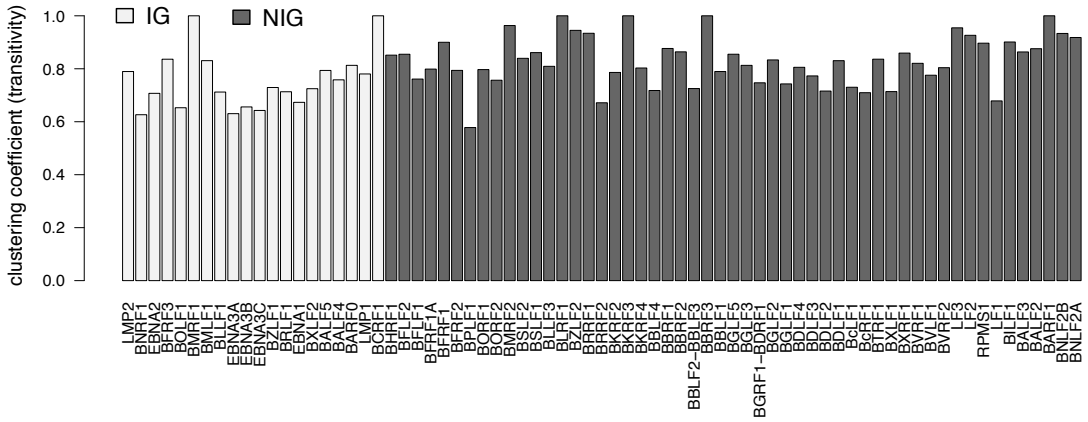
(A) Oberserved and expected number of links between nonsynonymous sites between different categories of genes (Chi-square test, p < 2.2e−16).



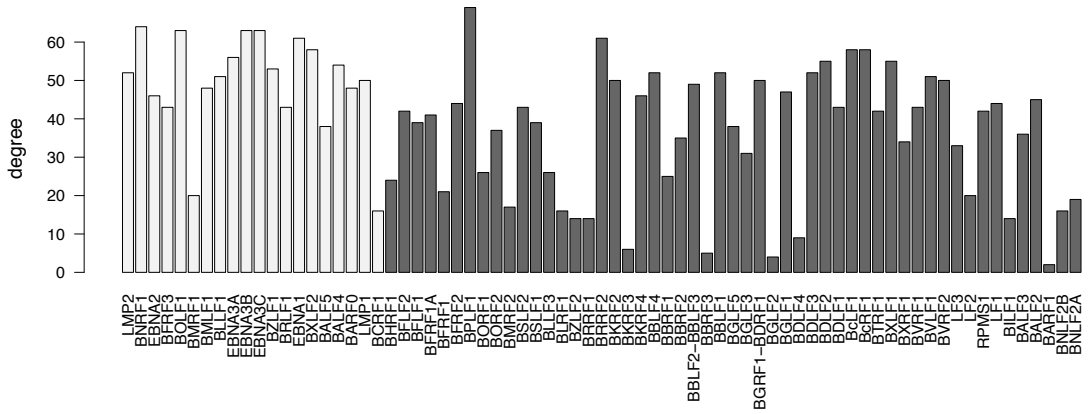
(B) Observed and expected number of links between nonsynonymous sites with a minimal distance of 1 kb between different categories of genes (Chi-square test, p < 2.2e−16).

FIGURE 4.7

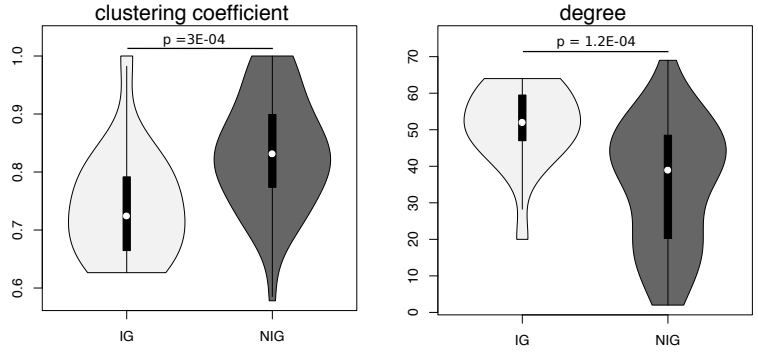
IG-NIG and NIG-NIG nodes (Mann-Whitney U test, p = $1.86e{-7}$ for IG-IG vs. NIG-NIG, and p = $3.55e{-5}$ for IG-IG vs. NIG-IG, respectively; supplemental figure B.9). The distance of genes towards each other does not have an effect on the linkage score between nodes (no significant difference between all combinations of inter-gene distance classes with Mann-Whitney U test, supplemental figure B.10).



(A) Bar plot of the clustering coefficient (transitivity) for every node in the network.



(B) Bar plot of the degree of every node in the gene network.



(C) Violin plots of the clustering coefficient (transitivity) and degree divided into the vertex types IG and NIG. There is a significant difference between the groups IG and NIG for both measures (Mann-Whitney U test).

FIGURE 4.8: Gene network properties.

Gene network properties and topology Figure 4.8a shows the local transitivity, or clustering coefficients, for every gene (node) of the network. The clustering coefficient C for a node i is defined as

$$C_i = \frac{2e}{k(k-1)}$$

where k is the degree of the node i (the number of edges with which it it is connected

to ther nodes, which equals the number of its neighbours) and e the number of edges between the k neighbours of i. In other words, it measures the ratio of the number of edges between the neighbours of i to the total possible number of such edges (Pavlopoulos et al., 2011). It is bounded by 0 and 1. The closer it is to 1, the more likely the network is to form clusters. In general, the clustering coefficients are fairly high for every node, indicating that the clusters are not separate but interconnected (because every node has many interconnected neighbours). However, nodes of type IG have a significantly lower clustering coefficient than those of type NIG (Mann-Whitney U test, $p = 3\text{e}{-4}$; figure 4.8c). The global clustering coefficient of the network is $C_g = 0.75$. This high overall value again underlines that the graph is quite dense. The actual density of the graph (ratio between the number of edges to the possible number of edges) is 0.54 and supports this notion.

The degree of every node is shown in figure 4.8b. Again, there is a significant difference between the degree of nodes belonging to the type IG and NIG, with NIG having in general fewer neighbours than IG (Mann-Whitney U test, $p = 1.2\text{e}{-4}$; figure 4.8c).

Comparing the two plots 4.8a and 4.8b, one node that sticks out is *BPLF1* of type NIG, which, interestingly, has the lowest clustering coefficient ($C_{BPLF1} = 0.58$) but the highest number of neighbours (degree) of all nodes ($k_{BPLF1} = 69$), i.e. it is linked to 69 of the 72 other nodes in the network. This means, many nodes are connected to it, but these neighbours are not necessarily all interconnected with each other. In terms of the network topology, this indicates, that *BPLF1* behaves a little bit like a hub in the network.

Another characteristic of a network is whether it is assortative or dissassortative. If nodes with a certain characteristic (for example a high degree of connectivity or a vertex label) have the tendency to be connected with other nodes of the same characteristic, it is called assortative. It is dissassortative, on the other hand, if high degree nodes have a tendency to be connected to low degree nodes. The measure of the assortativity coefficient is basically equivalent to the Pearson's correlation coefficient. The degree assortativity coefficient is $r_{degree} = -0.16$, and it is $r_{types} = -0.05$ if based on vertex label IG vs. NIG. The absolute of both values is not particularly high, meaning there is no clear tendency to be dissassortative or assortative in the way nodes are connected with each other.

To conclude this exploratory section, the network is fairly dense and strongly connected (i.e. there are no disconnected components).

Identifying biologically meaningful subnetworks and ranking of nodes The are several approaches to identify important subgraphs or nodes within the network:

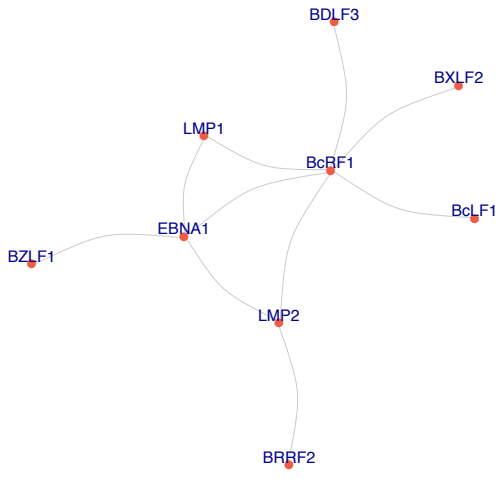
1. Community clustering on the graph to find one or several (biologically meaningfully) connected subgraphs. These are usually based on paths and walks through the graph and the optimisation of some modularity score;
2. Clustering based on the adjacency matrix of $n \times n$ nodes (n being the number of nodes in the network) where every entry between node i and j is the edge weight of the two nodes;
3. Ranking of nodes based on network centrality.

Various graph clustering algorithms implemented in **igraph** were applied, such as:

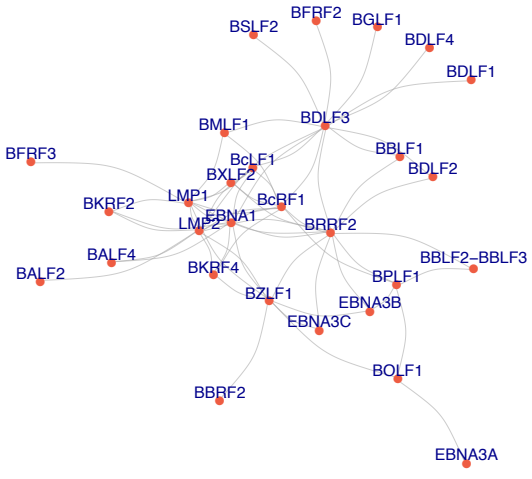
- Clustering based on *edge betweenness*, i.e. the modularity of the network, where a group of nodes is densely connected to themselves but sparsely connected to other modules. The modularity score of the network is negative, however, indicating that the network does not consist of modules. It identified 34 clusters, 32 of them consisting of a single node, and two clusters containing 6 and 34 nodes, respectively.
- The *walktrap* algorithm, which tries to find densely connected subgraphs via random walks. It identified nine clusters.
- The *fast greedy* algorithm tries to find subgraphs by optimising the modularity score. It identified four clusters.
- A last algorithm tries to find structure based on *propagating labels*. In the beginning, each node has a unique label, and at every step, each node adopts the label of the majority of the neighbours. In this gene network, this resulted in only one cluster.

These clustering methods yielded inconclusive results (supplemental figure B.11), i.e. identified clusters varied greatly in number and showed little overlap. This is probably due to the strongly connected nature of the network, as most of these algorithms are based in some way on the topology of the network. This approach was therefore discarded.

Figures 4.9a and 4.9b show the subgraphs of the network that are connected with edges with weights of A) top 1% and B) top 5% of the linkage score. The histogram shows that very high linkage scores are rare, and even within the top 5%, values range as low as 0.18.

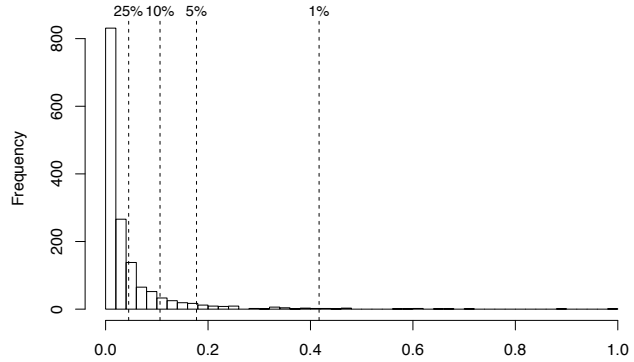


(A) Subgraph with a minimal linkage score threshold of 0.42 (top 1 %).



(B) Subgraph with a minimal linkage score threshold of 0.18 (top 5 %)

Figure continues on the next page.



(C) Histogram of linkage score values and cut offs of top percentiles.

FIGURE 4.9: Filtering of the gene network based on the edge weight (linkage score).

Figure 4.10 shows the hierarchical clustering approach based on the adjacency matrix. The distance matrix used for the clustering is based on the linkage score, i.e. the higher the linkage score between two genes, the smaller the distance between them. This is a similar approach (in this case of a highly connected network) of filtering based on the linkage score as in figure 4.9. Nodes belonging to IG are coloured orange, those belonging to NIG are coloured blue. Marked in grey are the nodes that show some weak clustering: There are two clusters that are successively being formed with increasing distance (i.e. decreasing linkage score) until they are joined. *LMP1* and *BcRF1* form the first cluster which increases in size until the small cluster made of *EBNA1* and *LMP2* is merged with it. Further decreasing the distance from there leads only to other nodes being successively added to this cluster. This indicates again the network is very interconnected and not highly structured, but this approach was able to identify one small cluster of particularly strongly linked genes.

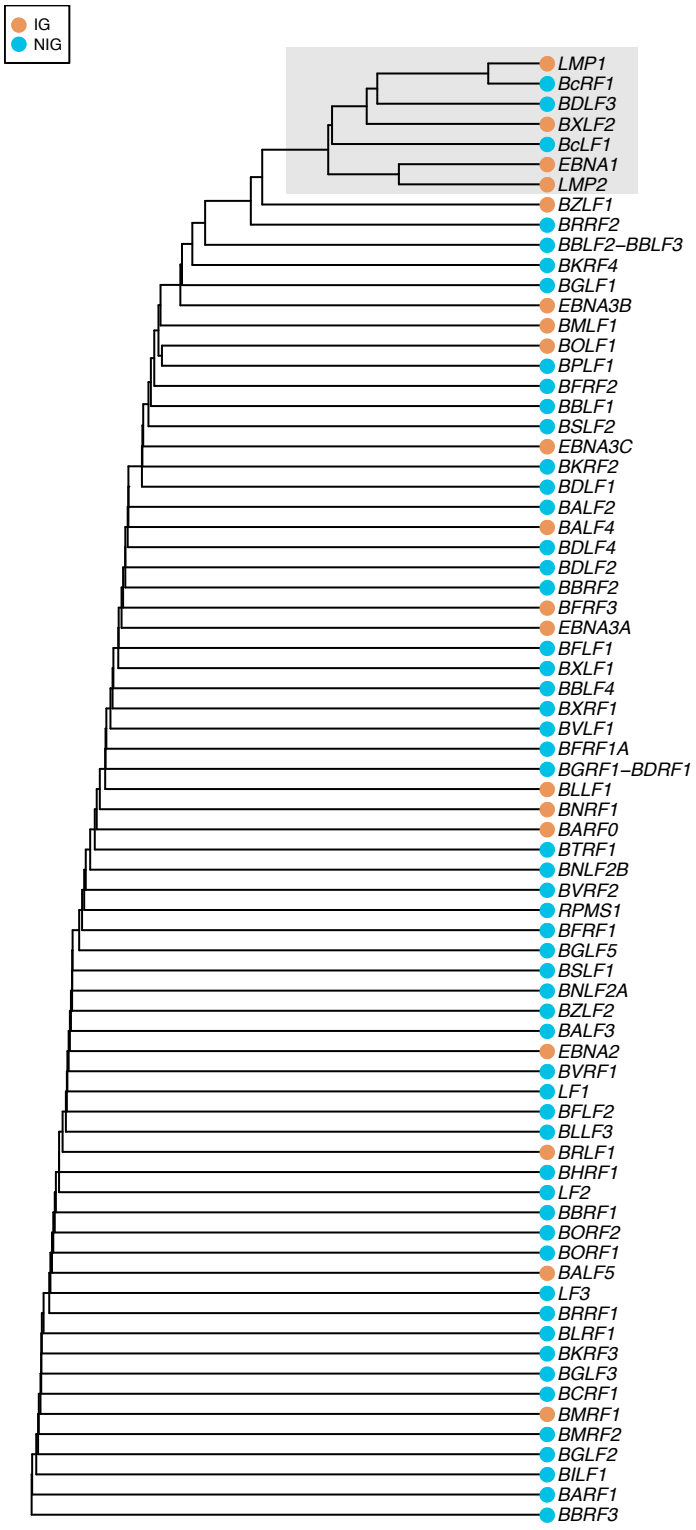


FIGURE 4.10: Hierarchical clustering of gene network. The distance is calculated based on their linkage score (edge weight), i.e. the higher the linkage score between two genes, the smaller the distance between them. Nodes belonging to IG are coloured orange, nodes belonging to NIG are coloured blue. Marked in grey are the two main clusters that are being joined at low distance already.

Alternatively, identifying the most important genes in a network can be done by ranking nodes based on their properties. Eigenvector centrality does this by measuring the influence of a node, i.e. a node's score is higher if it is connected to other high-scoring nodes. Figure 4.11 illustrates this. All genes are coloured based on their ranking within the network. Of the top 25 highest ranked genes, 13 belong to the group IG (in total there 19 IG nodes) (table 4.2). However, five more of these genes also appear in the IEDB database as antigens, but did not meet the rather conservative criterion of a minimum of two independent references.

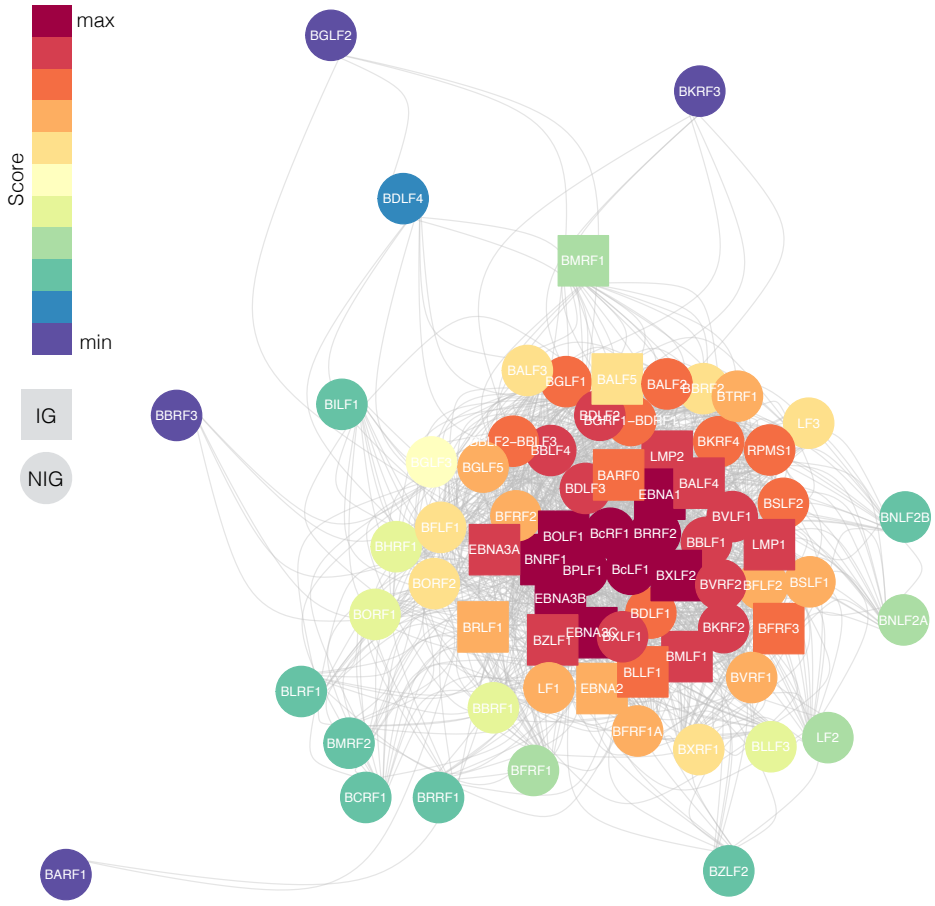


FIGURE 4.11: Gene network coloured by Eigenvector centrality.
Gene network coloured by Eigenvector centrality, with warm colours indicating higher and cooler colours lower scores, respectively. Square node symbols denote genes belonging to IG, circular nodes denote genes belonging to NIG.

Eigenvector rank	ORF	Protein	IG
1	BPLF1	Large tegument protein deneddylase	○
2	EBNA3B	Epstein-Barr nuclear antigen 4	●
3	BOLF1	Protein BOLF1	●
4	BNRF1	Major tegument protein	●
5	EBNA3C	Epstein-Barr nuclear antigen 6	●
6	EBNA1	Epstein-Barr nuclear antigen 1	●
7	BRRF2	Tegument protein	○
8	BcLF1	Major capsid protein	○
9	BXLF2	Envelope glycoprotein H	●
10	BcRF1	TBP-like protein	
11	BALF4	Envelope glycoprotein B	●
12	BDLF2	BDLF2 (Glycoprotein)	
13	BXLF1	Thymidine kinase	○
14	LMP2	Latent membrane protein 2	●
15	BBLF1	Tegument protein UL11 homolog	
16	BDLF3	BDLF3 (Glycoprotein)	
17	BZLF1	Transactivator protein	●
18	BVLF1	BVLF1	
19	BVRF2	Capsid scaffolding protein	○
20	EBNA3A	Epstein-Barr nuclear antigen 3	●
21	BKRF2	Envelope glycoprotein L	
22	BBLF4	DNA replication helicase	
23	LMP1	Latent membrane protein 1	●
24	BMLF1	mRNA export factor ICP27 homolog	●
25	BARF0	BARF0	●

TABLE 4.2: Most influential nodes in the network. Circles in the column labelled IG mark proteins for which an immune response has been reported, with filled circles fulfilling the criterium of having at least two references and empty circles having fewer than two.

4.2.3 Prediction of novel T cell epitopes for EBV

Figure 4.12 shows the results of a number of evolutionary analyses plotted against the genome map of EBV.

Selection

To test for positive and negative selection, the summary statistic method Tajima's D (Tajima, 1989) was used and results are shown in track A. Tajima's D uses information about the number of polymorphic sites within an alignment to infer the presence or absence of selection on a gene in comparison to a null model of neutral evolution. Under the null model, D is expected to equal zero. Negative D values result from an excess of rare alleles, i.e. a high number of segregating sites compared to a low number of pairwise

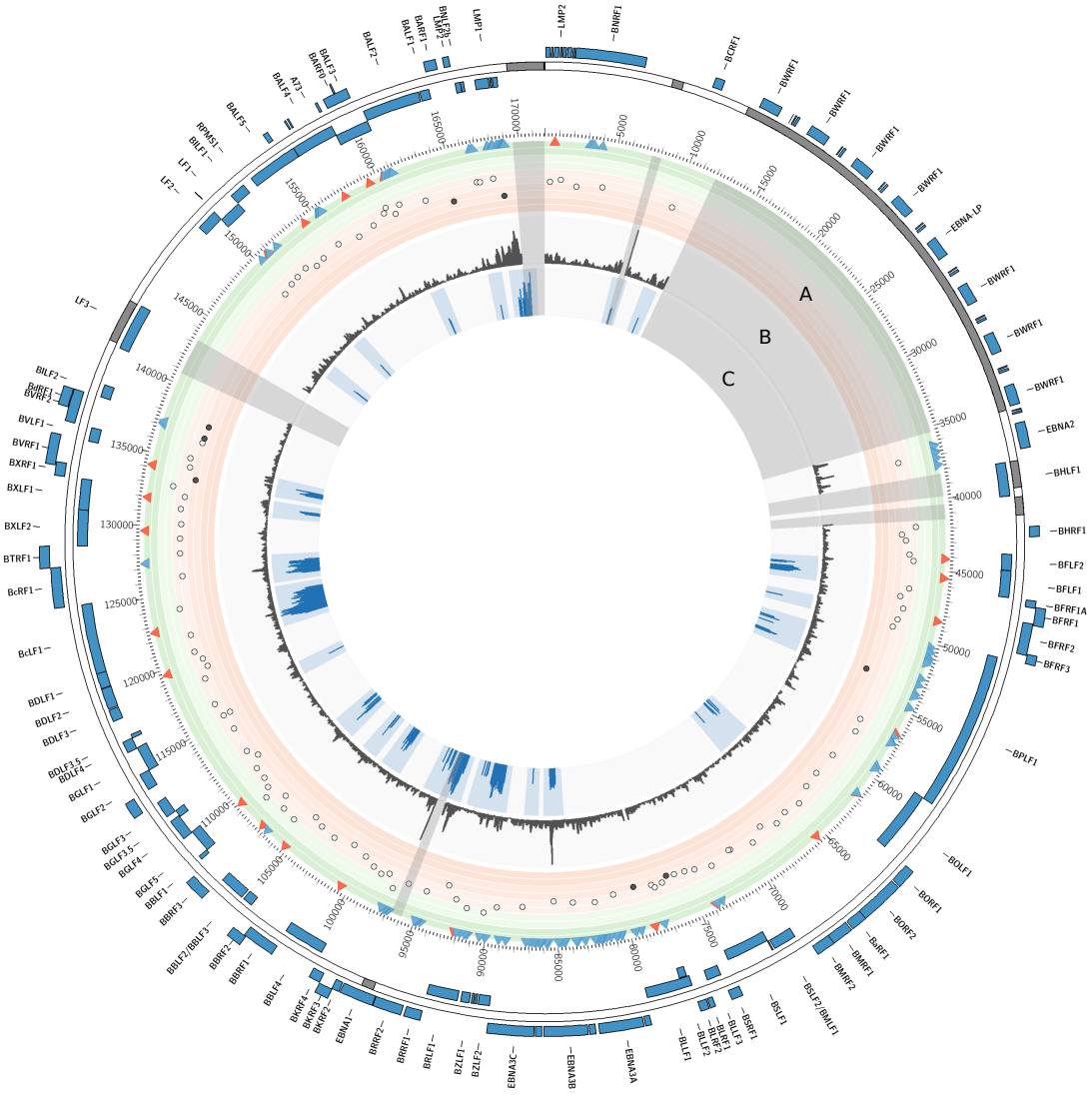


FIGURE 4.12: Circular EBV genome map showing the combined results of various evolutionary analyses. The outermost tracks marks the ORFs around the genome. The grey track inbetween and the grey shading mark the repeat regions that have been excluded from the analysis.
Track A: Red arrows show recombination breakpoints as detected with GARD. Blue arrows denote positive selected sites (PSS). Tajima's D values are plotted for every gene (or gene fragments in case of recombination). Red and green background refers to negative and positive values, respectively. Filled circles are tested significant with $p < 0.05$. Track B: Nucleotide diversity calculated in windows of 100 bp with step size of 1 bp. Track C: Genome-wide sliding window scan of LD in windows comprising 20 SNPs. Shown are only the windows of significant hotspots of local LD.

differences, indicating a) a recent bout of positive selection, b) purifying selection or c) a population expansion after a recent bottleneck. Positive D values, in contrast, result from an abundance of intermediate-frequency alleles, which can be indicative of a) balancing selection, b) a population structure or c) a decreasing population. Here, the majority of

ORFs (or fragments) had negative D values. The highest positive D value was found in *EBNA1* with 0.55, which is still a low value and not significant. Seven genes had significant, negative D values ($p < 0.05$, filled black circles in figure 4.12: *BALF1*, *BILF2*, *BLLF1*, *BLRF1*, *BPLF1*, *BVRF1* and *LMP1*). Of these seven genes, thee showed evidence of recombination (red arrows in track A mark the likely recombination breakpoint) and D values were significant for one of the recombining fragments, respectively (*BVRF1*, *BPLF1* and *BLLF1*). Given the data set, a recent population expansion is an unlikely scenario as EBV is spread worldwide and sequences are derived from different geographical regions rather than a small subset. Moreover, sampling dates vary over decades. This also makes it unlikely for the rare alleles to result from a recent bout of positive selection that have not yet reached fixation in the population. The most likely explanation is therefore that purifying selection is acting on these genes or gene fragments. All Tajima's D values are listed in supplemental table B.2.

Detection of specific sites within a gene that are under positive selection (positively selected sites, PSS) was performed using codeml from the paml package (Yang, 2007). Here, codon substitution models are used that allow the ratio of non-synonymous to synonymous substitutions to vary among sites. Specific codons under positive selection (PSS) were detected in 23 genes (figure 4.12, blue arrows in track A; supplemental table B.1) and are reported for both the model comparison M1a-M2a (nearly neutral evolution against positive selection) and M7-M8 (ω follows a beta distribution, with M8 additionally allowing for positive selection). With the exception of a few sites, all PSS were supported by both the simpler and robust model (M2) as well as the more sensitive and complex model (M8) (Yang, 2007) (supplemental table B.1).

A large number of PSS was found for the latency genes *EBNA3A-C*, *EBNA1* and *LMP1*. Most of them lie within known epitopes for CD4+ and CD8+ T cells and likely represent adaptations for immune escape.

Nulceotide diversity

Track B in figure 4.12 shows the nucleotide diversity in sliding windows of 100 bp. The ten ORFs with the highest diversity are in decreasing order *LMP1*, *EBNA1*, *BZLF1*, *BRRF2*, *EBNA3B*, *BDLF3*, *LMP2*, *EBNA3C* and *EBNA3A* (supplementary table B.2). Eight of these are immunogenic genes according to table 4.1.

The distribution of biallelic SNPs across the genome is shown in supplemental figure B.13. SNPs are not completely randomly distributed across the genome, instead there seem to be hotspots based on a Monte Carlo test as implemented in **adegenet** ($p = 0.001$).

Local linkage disequilibrium

A sliding window approach was used in order to detect local signatures of LD. The areas, which have been found to show significant high local linkage compared to the rest of the genome are shown in track C of figure 4.12. These areas fall within 29 ORFs as shown in table 4.3.

ORF	Protein	PSS
BALF3	Tripartite terminase subunit 1	•
BBLF1	Tegument protein UL11 homolog	
BBLF2-BBLF3	Helicase-primase subunit BBLF2/3	
BBLF4	DNA replication helicase	
BBRF1	Portal protein	
BBRF3	Envelope glycoprotein M	•
BcLF1	Major capsid protein	
BcRF1	TBP-like protein BcRF1	•
BDLF1	Triplex capsid protein 2	
BDLF4	Uncharacterized protein BDLF4	
BGLF1	Capsid vertex component 1	
BGLF5	Shutoff alkaline exonuclease	
BKRF2	Envelope glycoprotein L	
BKRF3	Uracil-DNA glycosylase	
BNLF2B	Uncharacterized protein BNLF2b	•
BORF1	Triplex capsid protein 1	
BORF2	Ribonucleoside-diphosphate reductase large subunit	
BPLF1	Large tegument protein deneddylase	•
BRLF1	Replication and transcription activator	•
BRRF1	Transcriptional activator BRRF1	
BRRF2	Tegument protein BRRF2	•
BVRF1	Capsid vertex component 2	
BXLF1	Thymidine kinase	
BZLF1	Trans-activator protein BZLF1	•
EBNA1	Epstein-Barr nuclear antigen 1	•
EBNA3B	Epstein-Barr nuclear antigen 4	•
EBNA3C	Epstein-Barr nuclear antigen 6	•
LF1	Uncharacterized LF1 protein	•
LMP1	Latent membrane protein 1	•

TABLE 4.3: ORFs and their encoded proteins that display local high LD and whether they contain any PSS.

Epitopes

The presence of local LD while being under positive selection and showing a high degree of diversity could be indicative of a protein being immunogenic. Maintaining diversity within epitopes helps the virus to evade the immune system. Many of these characteristics are being shown by known strong immunogenic proteins such as LMP1, the EBNAs and BZLF1. I therefore sought whether it is possible to identify proteins/genes, that show similar characteristics but are not known to be immunogenic.

The procedure of predicting T cell epitopes is described in detail in chapter 2. In short, protein sequences that contain positively selected sites (PSS) are being screened for peptides that can be presented by MHC (i.e. the "wild type" and mutant variant). A predicted peptide is being considered as a candidate for further analysis if it a) belongs to the top 1 % of predicted peptides (in total) and b) contains at least one PSS. The HLA-specific

binding properties are then compared between variants of the peptide based on their predicted IC50 value (derived from two different algorithms, ANN and SMM), which measures the peptide's predicted affinity, in other words whether a variation induces a strong effect in binding affinity to evade the immune system. Previous studies have suggested a IC50 value of 500 nM as a affinity threshold for HLA presentation associated with potential immunogenicity for HLA class I restricted epitopes (Sette et al., 1994). Generally speaking, IC50 values of <50 nM are considered high affinity, <500 nM intermediate affinity and <5000 nM low affinity. The majority of known epitopes have high or intermediate affinity. Peptides binding only with low affinity may exist, but there are no known T cell epitopes with IC50 values >5000 nM (*Immune Epitope Database*). A potential epitope has therefore been chosen if a variation leads to change in affinity.

"Positive control" First, however, I wanted to see, whether this approach works for known immunogenic genes and whether it is possible to predict epitopes that have been described experimentally. The prediction approach was applied to the latency proteins *EBNA1*, *EBNA2*, *EBNA3A*, *EBNA3C* and *LMP1* and then compared the prediction results to known epitopes listed in IEDB.

Table 4.4 shows the results for this. For 4 proteins, the same or similar epitopes were identified.

Gene	PSS	Pred. HLA	Pred. peptide	Exp. HLA	Exp. peptide
EBNA1	411	B*53:01	HPVGEADYF	B*53	HPVGEADYF
		B*35:01	HPVGEADYF	B*35:01	HPVGEADYFEY
		DRB3*01:01	PFFHPVGDADYFEYL	ND	RPFFHPVGDADYFEY
EBNA3A	333	B*08:01	FLRGRAYGI	B*08	FLRGRAYGL
	459	B*35:01	YPLHEQHGM	B*35:01	YPLHEQHGM
	655	DRB1*08:02	QVADVVRAPGVPAMQ	ND	QVADVVRAPGVPAMQF
EBNA3C	141	DRB1*03:01/ DRB3*02:02/ DRB3*01:01	ILCFVMAARQRLQDI	DR13	ILCFVMAARQRLQDI
LMP1	18	DRB1*09:01	PPRRPPLSSSLGLAL	DR9	GPPRRPPLSSSLGLALL

TABLE 4.4: Positive control for epitope prediction procedure. The table lists those genes and the respective PSS for which a peptide was predicted to be an epitope as well as the predicted HLA allele. It compares it to the experimentally confirmed peptide and its HLA restriction. The PSS position refers to the amino acid position in the protein.

Prediction of novel epitopes I then applied this procedure to three proteins: BcRF1, BRRF2 and BPLF1. *BcRF1* displays a certain degree of variability and contains highly locally linked sites. There are two sites under positive selection. Tajima's D is negative (-0.91) but not significant, indicating there is a slight excess of rare alleles. *BRRF2* has two PSS, shows the fourth highest nucleotide diversity and is also a local LD hotspot. Last, *BPLF1* has 21 PSS and one recombining fragment is under significant purifying selection ($D = -2.9$). The two selection tests are not conflicting, as positive selection can act on

specific sites, while over the whole gene, purifying selection can be the driving force of evolution. There's a hotspot of local linkage within the same fragment.

Interestingly, in the previous analysis of the gene network of linked, nonsynonymous sites, a significantly higher number of neighbours (degree of a node) was found for immunogenic genes than for nonimmunogenic genes (figure 4.8c). The ORFs encoding for the three selected proteins, which have been classified as NIG, also have high degrees compared to the other NIG nodes (figure 4.8b).

Gene	PSS	HLA	Start	End	Length	Peptide	IC50 ANN	IC50 SMM
BcRF1	738	A*03:01	729	738	10	ITLLLAHLRK	42	19.33
						ITLLLAHLRR	879	274.92
		A*03:01	730	738	9	TLLLAHLRK	44	66.34
						TLLLAHLRR	888	274.92
BRRF2	184	B*08:01	175	184	10	LVALRGHVQL	1126	454.59
						LVALRGHVQP	22899	3833.72
		DRB1*03:01	181	195	15	HVQLAYDARVLTPDF	160	30
						HVQPAYDARVLTPDF	1922	763.1
		DRB1*03:01	180	194	15	GHVQLAYDARVLTPD	162	17.5
						GHVQPAYDARVLTPD	1762	1372.8
	260	B*07:02	255	263	9	APETLRDYL	123	180.26
						APETLQDYL	2972	638.1
	323	B*07:02	321	329	9	RPRFSALPP	28	106.88
						RPQFSALPP	655	632.25
	323+325	B*07:02	321	329	9	RPRFSALPP	28	106.88
						RPQFLALPP	1693	1034.88
BPLF1	12	B*07:02	11	20	10	RPRGTGPVRG	217	122.13
						RTRGTGPVRG	18374	2740.44
		B*07:02	11	18	8	RPRGTGPV	66	462.39
						RTRGTGPV	908	21330.9
		B*07:02	11	19	9	RPRGTGPVR	100	42.84
						RTRGTGPVR	14526	3146.95
	796	B*08:01	788	796	9	LIRSRDRSA	26	83.67
						LIRSRDRSS	10066	83.76
	1535	A*01:01	1533	1541	9	FTDIETGPL	57	356.9
						FTEIETGPL	880	1440.62
	2895	B*07:02	2893	2901	9	APRPQKTQA	16	50.69
						APSPQKTQA	210	345.06
		B*07:02	2893	2902	10	APRPQKTQAQ	187	181.06
						APSPQKTQAQ	1997	1108.72
	2895+2897	B*07:02	2893	2901	9	APRPKKTQA	15	59.14
						APSPKKTQA	173	402.62
		B*07:02	2893	2902	10	APRPQKTQAQ	16	50.69
						APSPKKTQAQ	1561	924.31

TABLE 4.5: Candidate peptides from epitope prediction for both MHC I and II. Presented are the comparison of both aminoacid variants and their predicted IC50 value based on two algorithms (ANN and SMM). Start and end coordinates as well as the PSS position refer to the amino acid positions of the protein.

Based on our filtering criteria, a number of putative epitopes could be identified. The results are listed in table 4.5 and figure 4.13, which graphically illustrates how the changes in PSS highlighted in table 4.5 change the affinity of being presented by MHC. Note that

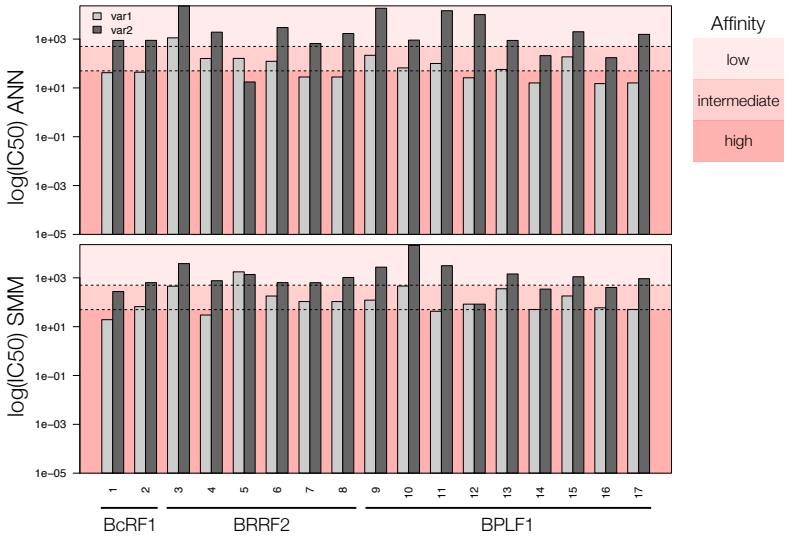


FIGURE 4.13: Predicted IC50 values (on a log scale) of the ANN (top) and SMM algorithm (bottom) of candidate peptides. Var1 (light grey) and Var2 (dark grey) refer to the changes at PSS as highlighted in table 4.5. The background colour indicates the affinity of this peptide being presented by an MHC molecule.

these are not distinct peptides. For *BcRF1*, two peptides for MHC I have been predicted. Both are highly similar and have a predicted specificity for HLA A*03:01, for which the variation from K to R decreases its binding affinity. For BRRF2, four peptides for MHC I and two for MHC II could be predicted. Again, the two predicted MHC II epitopes are highly overlapping. Moreover, there are two MHC I epitopes predicted containing two PSS that are very close together. Depending on the combination of variations, the binding affinity is further decreased. In the case of BPLF1, in total nine peptides that could be potentially be presented by MHC I molecules have been predicted, some of them being very similar or containing multiple PSS.

4.3 Discussion

4.3.1 Recombination and linkage disequilibrium

Recombination plays an important role in viral evolution as it is, beside mutations, a source of genetic diversity. By combining variants that occur on different genomes, it allows the introduction of new haplotypes.

Recombination is also a means to maintain the virus genome's integrity in the state of latency when it is constantly exposed to mutagenic agents. Employing the host cell's homologous recombination (HR) repair pathway has been proposed as a mechanism to circumvent this (Brown, 2014; Wilkinson and Weller, 2004). This is in concordance with the observation that DNA mismatch repair and recombination proteins are involved in herpesvirus replication (Wilkinson and Weller, 2004). Interestingly, α- and γ-herpesviruses

seem to utilise different HR initiating sequencing. In α-herpesviruses, inverted and tandem repeat sequences are the most prevalent, whereas it is specific short, GC-rich sequences in γ-herpesviruses (Brown, 2014).

It is therefore easy to imagine that in the case of co-infection, the employment of the host's HR-machinery might lead to recombinant strains.

The advancement in EBV whole genome sequencing has allowed for recombination to be observed (Palser et al., 2015; Kwok et al., 2014; Santpere et al., 2014). These studies have mostly looked at selected strains and showed recombination breakpoints between a number of genomes. Instead of focusing on specific strains, I wanted to look at recombination and its traces more comprehensively. Evidence for recombination breakpoints was found throughout the genome. This is in concordance with an *in silico* study that analysed the occurrence of potential recombination inducing sites relatively uniformly across the EBV genome (Brown, 2014).

Despite this, many variable sites were observed to be in linkage disequilibrium with each other, i.e. they always occur in the same strains together, even over long distances. This is surprising as recombination disrupts linkage between sites. Moreover, this seems to be different from other herpesviruses: In human cytomegalovirus (HCMV), a recent study found no LD in the majority of the genome due to unrestricted recombination, except for a few regions where recombination seems to be constrained in order to maintain diversity (Lassalle et al., 2016).

In general, LD can indicate a population structure, which might be due to geographic variation. However, evidence was still found for recombination occurring within the subset of sites in LD (supplemental figures B.1-B.6).

There are two possible scenarios: a) an ancient population structure that has only recently begun to recombine or b) this is a steady state maintained by selection for epistasis. Further work is necessary to completely answer this question. One can however speculate that an ancient structure being erased by recombination is the more parsimonious explanation in comparison to a high number of loci being under epistatic constraints, but both hypotheses are not mutually exclusive.

Two inter- and intra-recombining subpopulations

Population structure analysis revealed that the majority of the structure observed can be explained by dividing the set of sequences into two populations. One population seems to be mainly comprised of isolates from Asia, while the other population contains genomes from the rest of the world, i.e. isolates from Africa as well as "Western" countries such as Australia, the UK and the USA. This confirms previous findings where Asian type 1 sequences cluster quite distinctly from other genomes in a PCA (Palser et al., 2015). There are also a few admixed isolates, i.e. where the genome has been partially assigned to both populations, which would suggest recombination events between them. One sequence from the UK, LN824142, seems to be belong to the Asian population rather than the other one. However, the geographic label of the isolates is based on where they have been isolated and is not necessarily the actual origin of the virus. It is easily

imaginable that a sequence isolated in the UK or Australia, countries with a mixed ethnic population, could originally be an Asian strain, as primary infection often occurs through close family members in early age. This highlights a big drawback of this sort of analysis: More information about where virus actually comes from, the ethnicity of the donor etc. would be needed.

Recombination is also occuring within these subpopulations, as evidence was found for recombination within the subset of sites in LD (and by further filtering based on the class of sites, such as only nonsynonymous sites in LD, or a p-value threshold, supplemental figures B.1-B.6). In fact, database sequences from the dataset used have been described as recombinants. HKNPC2, for example, seems to be a recombinant of HKNPC7 and -9 (Kwok et al., 2014), all of them isolates from Hong Kong that have been clearly assigned to the Asian population (named by their accession number in figure 4.5 KF992564, KF992569 and KF992571, respectively). Similarly, LN824224 has partially high similarity to four other southeast Asian isolates (Palser et al., 2015).

Interestingly, for admixed individuals the majority of the site data set is assigned to the Asian population when being restricted to only nonsynonymous sites in LD (membership coefficients around 0.7 or higher, figure 4.5, panel C). This is also true for isolates, which, when considering all sites, have only a very small percentage of the genome assigned to the Asian population (membership coefficients of 0.2-0.5). This opens the possibility that there might be functional linkage, for example because of constraints due to protein-protein interactions (PPI), or because of immune-selection.

There are two publications on the EBV interactome, Calderwood et al., 2007 and Fossum et al., 2009. Based on their data, the EBV PPI networks of selected nodes identified via hierarchical clustering (figure 4.10) and centrality (figure 4.11) from the gene network are shown in supplemental figure B.12. The hierarchical clustering approach is based on the nodes' linkage score, i.e. it represents the most strongly linked nodes. However, there is only one known interaction between LMP1 and the major capsid protein (*BcLF1*) (figure B.12a), which is only a low confidence interaction (Calderwood et al., 2007) and makes little sense in the virus' life cycle. The PPI network of the centrality-ranked nodes (table 4.2, figure B.12b) shows more possible protein interactions. Some of them are low confidence interactions, some of them are questionable regarding their localisation and their respective roles in different aspects of the life cycle. For example, *BRRF2* is a tegument protein, whereas its potential interaction partner *LMP2A* is a host membrane bound protein expressed during latency mimicking antigen-independent B cell receptor signaling (Kang and Kieff, 2015). Similarly, *BDLF3* is a host membrane glycoprotein responsible for immune evasion through MHC internalisation during late lytic cycle (Quinn et al., 2016) whereas *EBNA3A* is active during latency in the nucleus (Kang and Kieff, 2015). Other proteins are not further characterised in their function, such as *BARF0*. However, some of the interactions make a little more sense: *BPLF1* which plays among other functions a role in the envelopment of the capsid interacts with both *BALF4*, an envelope glycoprotein, and *BOLF1*, a tegument protein possibly involved virion assembly; *BNRF1* and *BBLF1* are both tegument proteins.

While there are a number of interactions reported, the PPI data has to be considered with care. It is based on Yeast two hybrid (Y2H) screenings, a technique with a number of disadvantages. For once, a false positive rate of up to 70% has been reported in Y2H data sets (Deane et al., 2002). This can be because of non-specific binding due to unusually high protein expression from plasmids. Another reason is that proteins might theoretically be able to bind each other, but this does not necessarily occur under natural conditions, because they are expressed at different time points in the viral life cycle (structural proteins during late lytic cycle versus latent genes, e.g. in the case of *BcLF1* and *LMP1*) or they are localised in different compartments of the cell.

Results of the analysis of linked genes supports the hypothesis of HLA adaptivity, as I found that immunogenic genes are more often linked with each other than would be expected by chance. This refers to nonsynonymous sites, i.e. sites on which selection can act. Moreover, 13 of the 25 most influential nodes in the gene network are immunogenic genes. In addition, three genes for which epitopes were predicted *in silico* (table 4.5) are also included in this list.

One could therefore hypothesise, that it is beneficial for the virus to retain certain combinations of polymorphisms in immunogenic genes. This might be, because the virus has adapted to HLA alleles common in the subpopulation it is circulating in. This is in concordance with a model of non-overlapping combinations of epitope regions, that are being held in LD despite genetic exchange via recombination between pathogens in other parts of the genome (Gupta et al., 1996). Nevertheless, protein-protein interactions cannot be excluded either. In fact, both options are not mutually exclusive.

The whole matrix of sites in LD is vast, comprising of approximately 89,000 pairs of loci in LD. The major difficulty lies in disentangling the noise (e.g. sites in LD due to hitch-hiking) from the informative sites. Here, I included in the analysis all sites in LD with a corrected p-value of 0.05. This value might have been too lenient, and a stricter filtering to begin with could help reducing the noise in the data, highlighting more relevant linkages. Another idea could be to weight linked pairs differently depending on their distance. As observation of linkage of proximal sites is more likely, those linkages might have a lower weight than long range linkages.

Moreover, studying more closely the admixed individuals is of interest, as these are the ones in which sites have been kept in LD on the amino acid level despite recombination between the two subpopulations.

Additionally, it could be of interest to focus on specific interaction pairs to address specific biological questions. For instance, proteins known to interact with each other (EBNA2 and the EBNA3-family, for example) would be a good target to specifically see which sites are linked in which strains, in which domains these interactions fall, and how they might affect function.

Interestingly, a recent paper (Chiara et al., 2016) has been published that also investigated EBV's population structure on a slightly bigger, but largely overlapping data set

(including type 2). They found ten subpopulations of EBV, eight of them containing isolates that could be assigned with low admixture that showed correlation with the geographic origin. Their findings seem reasonable also in the light of recent migration and admixture.

Given how different Asian strains are from the rest of the world, it might be possible that there is a less distinct population structure present within the non-Asian genomes that was missed by *structure*. But when restricting the data set to non-Asian genomes, it was not possible to identify a fitting number of k with Evanno's method (Evanno, Regnaut, and Goudet, 2005), which is indicative of much stronger admixture of genomes.

Another obvious reason for the different findings is that the data here were analysed for a maximum number of populations of $k = 10$, thereby underestimating the number of populations. However, repeating our analysis for larger number of k (ranging from 0 to 20), $k = 2$ was still found to be the best fitting number of populations (Evanno, Regnaut, and Goudet, 2005). It is not clear from the paper by Chiara et al., 2016 which models implemented in *structure* were used and how the number of populations was inferred.

4.3.2 Epitope prediction

Based on the whole genome analysis of various EBV strains I was able to identify areas of the genome were hotspots of local LD correlated with higher nucleotide diversity. Recombination and mutation rates have opposite effects on nucleotide diversity, the former decreasing it while the latter increases it. However, low diversity regions can also be the result of strong purifying selection. The concomitant occurrence of high LD and high nucleotide diversity indicates that recombination occurs less frequently in these areas. In other words, there might be constraints on recombination in order to maintain diversity. One possible explanation for viruses is that of immune evasion.

These characteristics are being shown by genes that are known to code for strong immunogenic targets, such as the latency genes and *BZLF1*. Based on the results of applying the epitope prediction procedure to known epitopes, I was prompted to apply this approach to previously unknown or little studied immunogenic genes, which show similar characteristics.

By also taking into account the results of selection analysis, I was able to predict a number of peptides in the genes of *BcRF1*, *BPLF1* and *BRRF2* to be potential targets of immune recognition. It has to be noted that this analysis was restricted to HLA alleles common in Europe.

Little is known about *BcRF1* function, but it has recently been shown to form a complex with the TATT motif, which is often present in γ-herpesviruses instead of the more common TATA-box (Gruffat et al., 2012) in the promotors of late genes. The protein is involved in a viral complex with five more proteins (encoded by *BDLF3.5*, *BDLF4*, *BVLF1*, *BGLF3*, *BFRF3*) that is responsible and essential for the initiation of late viral gene transcription (Aubry et al., 2014). This mechanism of late viral transcription is conserved

within β- and γ-herpesviruses (Aubry et al., 2014) and is necessary for productive infection. To date, no epitopes have been described for this protein.

BPLF1 is the largest ORF in the EBV genome. It is expressed in lytic cycle and encodes the large tegument protein. Additionally to its Deneddylase activity, which contributes to prevention of S phase progression of the host cell, it displays deubiquitinase (DUB) activity (van Gent et al., 2014). By suppressing TLF- and TRAF6-mediated activiation of NF-κB, it plays a role in innate immune evasion (van Gent et al., 2014; Saito et al., 2013). The DUB catalytic domain can be found in the N-terminal part of the protein between sites 1-269. Three very similar peptides that have been predicted as epitopes fall into this area of the protein. The majority of the gene has been positively tested with Tajima's D, indicating strong purifying selection selection due to functional constraints. However, many sites have been detected to be under positive selection.

Finally, the late gene *BRRF2* encodes another tegument protein, but is functionally not well described. It is likely located in the tegument (Johannsen et al., 2004) but has also been detected in the cytoplasm of cells during lytic cycle (Watanabe et al., 2015). During analysis, no T cell epitope was known for this protein; however, elevated antibody titres against *BRRF2* in multiple sclerosis patients have been described (Cepok et al., 2005). I was able to predict a number of potential epitopes. One of them (RPRFSALPP, position 321-329) contains two PSS very close together, and the additional mutation further decreases the affinity to bind HLA B*07:02. Interestingly, the IEDB has been updated since this work has been done and in fact, a very similar peptide (VPRPRFSAL, position 319-327) has been previously determined similarly by *in silico* prediction and further its binding affinity for B*07 validated *in vitro*, although it eclicited only a weak IFN-γ response in primary PBMCs from EBV^+ donors (Turčanová and Höllsberg, 2004).

It would be of great interest to conduct a similar study in order to validate the binding specificity *in vitro* and also to confirm whether a response in primary PBMCs of EBV^+ positive donors can be detected. This is also why the selection of candidate peptides has been restricted to those with binding specificities for HLA alleles common in Caucasians.

A possible drawback of our approach is the restriction to sites under positive selection. There are other regions in the genome that show a reasonable amount of variation while being strongly in local LD, that would also functionally make sense to be immunogenic. One such gene is *BcLF1* which encodes the major capsid protein. It contains a very big hotspot of local LD. Moreover, there are three epitopes described experimentally in the IEDB, though they have not been confirmed by further studies.

In addition to that, some epitopes in EBV are highly conserved (Duraiswamy et al., 2003; Chiu et al., 2014; Palser et al., 2015), and not all epitopes are necessarily under positive selection. Due to the dual nature of EBV's lifestyle, one can imagine that some epitopes in latency genes, in particular those driving the proliferation of B cells (*EBNA2*, *EBNA3A*, and *EBNA3C*), are partly under purifying selection. Hiding too efficiently from the immune system at this critical stage would likely increase the cancer rate by promoting unrestricted proliferations, thereby shortening the lifespan of the host and in consequence the time available to EBV for transmission.

Another disadvantage is that the quantitative estimations of IC50 values from the prediction tools deviate systematically from experimental IC50 values (*Immune Epitope Database*). The operators of IEDB are currently undertaking an evaluation of the correlation between predicted IC50 values and the antigenicity of peptides.

Chapter 5

Whole genome sequencing of EBV from various clinical settings

5.1 Introduction

EBV plays a role in a large number of malignancies. Here, I used whole genome sequencing in combination with targeted enrichment to retrieve and analyse EBV whole genome sequences from various clinical settings, including: immunodeficiency after transplantation/Post transplant lymphoproliferative disorders (PTLD), infectious mononucleosis (IM), and nasopharyngeal carcinoma (NPC).

Posttransplant lymphoproliferative disorders Posttransplant lymphoproliferative disorders (PTLD) is a collective term for haematopoeitic lesions that occur in patients having undergone transplantation. The majority are of B cell origin, but 7-15 % are of NK or T cell origin (Swerdlow, 2007). PTLD is highly associated with EBV infection with around 60-80 % of PTLD being EBV^+ (Capello, Rossi, and Gaidano, 2005). PTLD can be classified into two groups: 1) early, polyclonal lesions and 2) monoclonal lesions that include polymorphic and monomorphic PTLD (Swerdlow et al., 2008).

The loss of immune control over EBV replication is thought to be one of the main causes for the lymphoproliferation. This is supported by the correlation between PTLD incidence and immunosuppressive dose (Juvonen et al., 2003) and the increase of EBV infected B cells in blood and tissue (Gulley and Tang, 2010). Consequently, a key component of PTLD treatment is the withdrawal of immunosuppression (Odumade, Hogquist, and Balfour, 2011).

In this chapter, I will present the results of sequencing and analysis of EBV genomes from tumours and blood of immunocompromised paediatric patients, to answer the question whether the virus found in tumour and blood is the same. Further questions relating to this are: Is it possible to detect multiple infections in blood or tumour? How does the virus change over time in immunosuppressed patients?

Infectious mononucleosis Upon primary infection, EBV can cause infectious mononucleosis (IM) with an incubation period of approximately six weeks. The symptoms of IM include a sore throat, cervical lymph node enlargement, fever and fatigue. While most symptoms last only between two and four weeks, fatigue can last for months (Balfour,

Dunmire, and Hogquist, 2015). Diagnosis of IM cannot be determined solely based on clinical manifestations, but has to be confirmed, for example with a heterophile antibody test (Basson and Sharp, 1969).

Immunoglobulin M antibodies targeting the viral capsid antigen (VCA) are being produced during the acute (0-3 weeks after onset) and sub-acute phase (4 weeks-3 months) by around 75 % of patients (Balfour et al., 2013), but diminishes during convalescence (4-6 months). After the onset, IgG antibodies targeting VCA are produced by all patients (Balfour et al., 2013), whereas IgG antibodies against EBNA1 develop only during convalescence (Balfour, Dunmire, and Hogquist, 2015). Profiles of these EBV-specific antibodies are therefore being used to determine the phase of EBV infection.

During acute illness, viral loads are high both in the blood and saliva compartment. The peak of viral load in serum and PBMCs occurs in the first week after onset of illness, and decreases afterwards. The kinetics of this decrease, however, can be very different from patient to patient (Berger et al., 2001). In saliva, on the other hand, EBV DNA can be detected for at least 6 months after onset of illness. This indicates a consistent infectivity of saliva even after symptoms have declined (Fafi-Kremer et al., 2005).

In this chapter, I will report on the sequencing and analysis of longitudinal samples of paediatric IM patients from Japan. Questions behind this analysis are the following: Is there a change over time observable within the genomes? Is the primary infection due to a single virus variant or can we detect multiple infections (over time)? If we observe changes in the genomes, where do they occur and do they correlate with changes in viral load?

Nasopharyngeal carcinoma Nasopharyngeal carcinoma (NPC) is, globally seen, a rare tumour. However, prevalence rates vary greatly geographically. For example, there is less than 1 case in 100,000 populations in North America and Europe. In Southern China, South East Asia and North Africa, however, rates are higher. In Southern China, for example, the annual age standardised incidence rate is 20-30 cases per 100,000 populations in men, and 8-15 cases per 100,000 populations in women (WHO, 2014).

NPC is a cancer of epithelial origin, linked with latent EBV infection and often infiltrated by lymphocytes. Especially in endemic regions, tumours are always EBV^+ (Chua et al., 2016).

Here, I did not sequence additional samples from the NPC setting. However, there are multiple, mostly Asian, NPC-derived EBV genomes available in the Genbank database. In fact, the majority of EBV genomes from Asia are NPC-derived. Therefore I sought to compare the publicly available Asian NPC genomes to the IM-derived genomes from Japan as representatives of Asian sequences from a NPC non-endemic region where incidence rates are low and stable (Kimura et al., 2011). To date, comparison between NPC-endemic and non-endemic regions has been restricted to specific ORFs (Sandvej, Zhou, and Hamilton-Dutoit, 2000).

5.2 Results

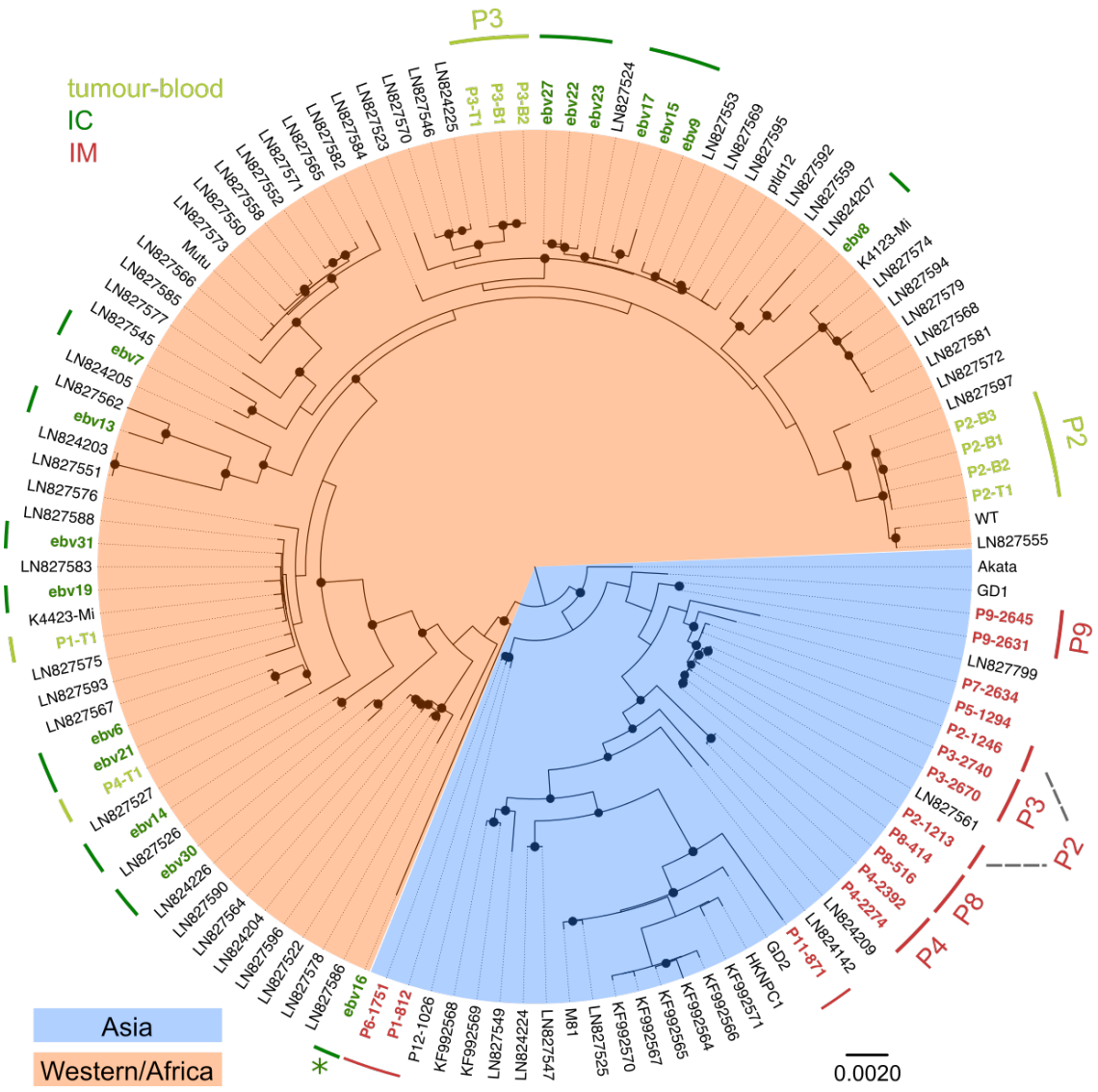


FIGURE 5.1: ML tree of published type 1 genomes including sequences from this chapter. Black circles indicate bootstrap values >90. Highlighted in blue: Asian cluster. Highlighted in orange: "Western"/African cluster (see chapter 4). Marked in light green: paired tumour and blood samples from paediatric patients in the UK; dark green: immunocompromised (IC) paediatric patients from the UK; red: paediatric IM samples from Japan. *: sequence classified as type 2. Specifically marked with a "Px" are patient for whom multiple samples are available (from tumour-blood and IM data sets).

Figure 5.1 shows a phylogenetic tree of the published sequences described in chapter 4 as well as all novel genomes generated in this project that will be individually highlighted and discussed in the following sections.

Note that not all nodes in this tree are well supported (i.e. bootstrap values >90, as indicated by a black circle in the tree). This is likely due to recombination which affects

tree topology and has been discussed earlier (chapter 4). However, towards the root and towards the leaves of the tree, nodes are usually well enough supported, especially in the cases where multiple isolates from the same patient were available, so that isolates can be put into context with each other and roughly to other published genomes.

5.2.1 Paired blood and tumour samples of immunocompromised paediatric patients

DNA extracts from whole blood and tumour biopsies from paediatric solid organ transplant recipients were enriched for EBV specific sequences using the 200 ng SureSelect protocol on the Bravo automation system and sequenced on the Illumina NextSeq platform. Table 5.1 lists sample and sequencing information. The blood sample of patient 1 was excluded, as coverage and depth was poor (33 % of the genome covered with an average depth of 5). Similarly, the blood sample of patient 4 was excluded as it failed library preparation. In the further analysis, only samples marked by a • in table 5.1 are considered ($n = 9$).

Genomes were quality checked and *de novo* assembled using pipeline 2 (see chapter 2).

Figure C.1 shows the coverage plots of retained samples after removal of duplicates. Blood samples had lower depths across the genome, while the tumour samples had a higher data output. This was expected as blood samples have much lower viral loads.

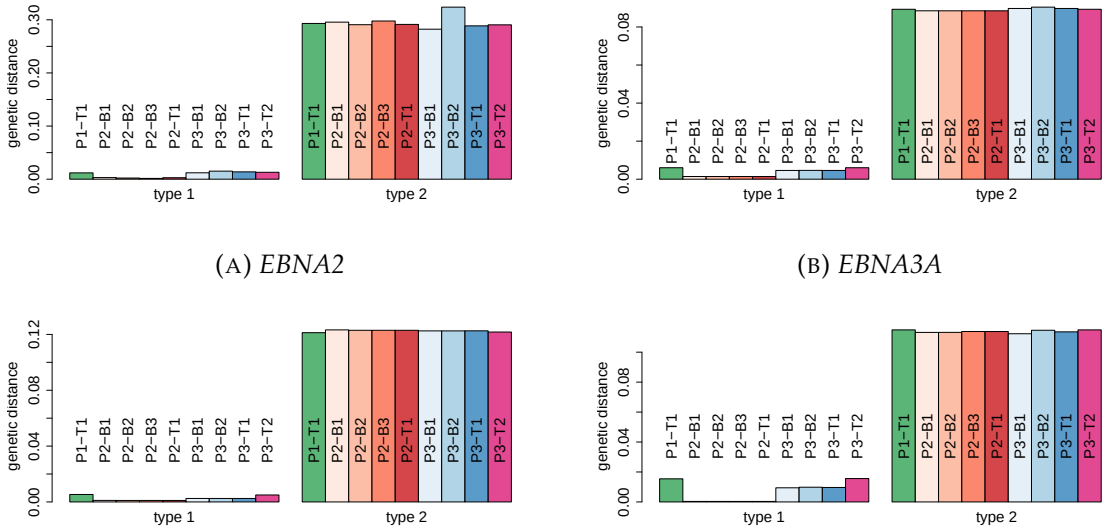
In consequence, we have paired tumour and blood samples for two patients with good coverage, both in depth and across the genome, as well as two additional tumour samples.

Patient	Sample	Date	Source	Viral load [copies/ml]	Read pairs	OTR %	Depth	Cov	
1	P1-B1	26/05/2015	blood	66,000	4,978,502	0.8	3	33	
	P1-T1	29/08/2015	tumour	11,229,000	5,973,819	79.1	4892	100	•
2	P2-B1	04/10/2012	blood	22,000	4,827,991	2.1	26	99	•
	P2-B2	18/11/2015	blood	1,707,000	5,936,426	2.3	30	97	•
	P2-B3	18/01/2016	blood	4,534,000	5,227,303	3.6	30	99	•
	P2-T1	02/10/2012	tumour	ND	5,718,349	77.2	4508	100	•
3	P3-B1	03/11/2011	blood	22,000	3,924,866	2.2	45	94	•
	P3-B2	03/09/2014	blood	952,000	7,450,997	2.2	16	100	•
	P3-T1	09/09/2011	tumour	ND	5,483,779	66.9	2004	100	•
4	P4-B1	15/03/2003	blood	>1,000,000	-	-	-	-	
	P4-T1	26/03/2003	tumour	ND	7,996,268	81.7	5885	100	•

TABLE 5.1: Sample and sequencing information for paired tumour and blood samples. OTR: On target reads in percent. Cov: Percentage of the genome that could be covered (repeats excluded). Depth: Average read depth after removal of duplicate reads. All samples that have been included for further analyses are marked by a • .

Description of consensus sequences

In order to determine the type of EBV patients are infected with, the distances of the typing genes *EBNA2* and *-3* genes to the type 1 and type 2 reference sequences (WT and AG876) were calculated. In all samples, the genetic distances were greatest to type 2, indicating all patients carry EBV type 1 (figure 5.2).



(A) *EBNA2*

(B) *EBNA3A*

(C) *EBNA3B*

(D) *EBNA3C*

FIGURE 5.2: Genetic distance (K80 model) of *EBNA2* and *-3* genes of paired tumour and blood samples to type 1 (WT) and type 2 (AG876) reference sequences.

Figure 5.1 depicts by means of a phylogenetic tree how the isolates fall into context with each other and other published genomes.

All paired tumour and blood samples (coloured in light green) lie within the cluster of Western/African sequences (tree highlighted in orange, see chapter 4). Samples from the same patient (of patient 2 and 3) cluster closely together and are very similar to each other as indicated by the short branch lengths. The tumour and first blood samples were taken close together ($\Delta t_{P2} = 2$ d and $\Delta t_{P3} = 55$ d), while follow up blood samples were taken several years later (table 5.1). Despite this, variation over time is very limited. No consensus level changes in patient 2 could be detected. For patient 3, there are three SNP differences (table 5.2). Two of the SNPs lead to an amino acid change: V250I occurs in the *BORF1*-encoded triplex capsid protein and I343M in the *BGLF1*-encoded capsid vertex component 1 (CVC1). Neither fall within known epitopes or known functional domains of the proteins.

The blood samples at both time points have mixed allele frequencies at these loci as observed in the pileup files of the assembly, even if the consensus base call is not an ambiguous one, while the tumour sample is homogenous at these positions. This suggests that either the mutations were acquired by some viruses in the blood after the

tumour was established, or that the virus present in the malignant B cell which formed the tumour did not carry these variations.

Therefore, the next step was to look more comprehensively at the intrahost variation in these patients and to compare the diversity found in blood versus tumour.

Sample	63696	115057	137837
P3-T1	G	A	A
P3-B1	R	C	G
P3-B2	A	A	R
nonsyn	•	•	
ORF	*BORF1*	*BGLF1*	

TABLE 5.2: SNPs in longitudinal data of paired blood and tumour samples of patient 3 on the consensus level. The isolates are ordered by sample date. If variants variants occurred at a frequency level of 0.5, ambiguity codes are introduced (R= G/A), otherwise the base with a frequency >0.5 is called. N denotes missing data. Positions are WT coordinates.

Description of intrahost variation

As stated previously, the average read depth varies greatly between tumour- and blood-derived samples, as the tumour has likely a much higher viral load (although the actual viral loads have not been determined for these particular samples). In order to reduce the impact of depth when comparing diversity of intrahost samples from different compartments of the same patient and across patients, the tumour samples have been subsampled to 34k mapping reads in order to achieve a comparable average depth of 30.

The resulting average depth across all samples is fairly low (on average 29.5). Cut-off values were therefore chosen for minimal depth as 20 and the minimal frequency as 20 %, as these settings allowed the condition of having 2 independent reads per strand to support the variant. For further parameters see chapter 2.

The fraction of the (subsampled) genomes with depth ≥ 20 was as this:
P1-T1: 0.6, P2-B1: 0.6, P1-B2: 0.7, P2-B3: 0.7, P2-T1: 0.7, P3-B1: 0.8, P3-B2: 0.2, P3-T1: 0.7, P4-T1: 0.7, without excluding the repeat regions in this fraction. In some cases, only a small portion of the genome fulfils this basic conditions for variant calling, indicating that possible variation even at medium frequency is likely not going to be detected.

Table 5.3 lists the minority variants found in this data set. The subsampled tumour samples is at this frequency level (20 %) conserved even on the minority level, as no variants in those samples were detected. For patient 2, three variants have been found in the blood, two in the first sample (P2-B1), and another, different one in the second sample (P2-B2), all of them around a frequency of 21-22 %. For sample P2-B1, these two variants fall within the ORF *BLLF1*, which encodes the envelope glycoprotein gp350, and lead both to amino acid changes at two consecutive positions in the protein (S508I and P509L).

	Pos	Ref	Var	Cov	Read1	Read2	Freq	ORF	nonsyn
P2-B1	78339	G	A	27	20	6	22.22	BLLF1	P509L
	78342	G	A	28	21	6	21.43	BLLF1	S508I
P2-B2	54559	G	C	24	18	5	20.83		
P3-B1	57474	G	A	43	33	10	23.26		
	115271	C	A	54	33	21	38.89	BGLF1	S272F
	138051	G	A	43	29	14	32.56	BILF2	G62V
P4-T1*	113928	G	A	7468	7005	461	6.17		

TABLE 5.3: Minority variants of samples from immunocompromised children with PTLD. Pos: WT position. Ref: Consensus base. Var: Minor variant base. Cov: Read depth at this position. Read1/2: Number of reads supporting the consensus or variant base, respectively. Freq: Minor variant frequency. ORF: Open reading frame. nonsyn: Amino acid change from consensus to variant at the respective protein position.
The tumour sample marked by a * is not subsampled.

For patient 3, there are three minor variants found in the first blood sample (P3-B1), two of them leading to nonsynonymous changes in the proteins encoded by *BGLF1* and *BILF2*, respectively. *BGLF1* encodes the capsid vertex component 1 (CVC1) and *BILF2* codes for a predicted glycoprotein.

While epitopes have been described for gp350 and CVC1, the observed minor variants do not fall into them (*Immune Epitope Database*) or for that matter in any functional domain.

The position 63696 which was variable on the consensus level, was not reported as a minority variant as the coverage was too low at this position. Similarly, the other two consensus level positions are not reported even if they look heterogeneous when inspecting the mapping manually; however, a combination of low coverage and low mapping quality of some reads might have resulted in their exclusion.

Finally, the tumour derived isolates were analysed without subsampling in order to assess how clonal the viral population in the tumour truly is, as the greater depth allows calling of even low frequency variants. Variants were called with a minimal coverage of 80 and a minimal frequency of 5 % in order to fulfil the condition of having at least 2 independent reads per strand supporting the variant. However, even when using the complete read data, the majority of samples did not show any minority variants at this low frequency level. Only one silent mutation at a level of 6 % was detected in one sample (see table 5.3, sample P4-T1*).

Taken together this points towards very low diversity in PTLD patients, both between compartments as well as over time. Variability in the tumours is except for one position absent, whereas a few minor variants could be detected in the blood. However, it is feasible that a lot of the variability present in blood remained unobserved, as these isolates have generally very low read depth. Variants could only be called reliably at a medium frequency due to this restriction. This was even further limited by some areas not fulfilling the variant calling settings.

5.2.2 Samples from immunocompromised children with chronically high viral load

In addition to the seven successfully sequenced samples from chapter 3, further 16 samples were sequenced. All of these are blood samples from paediatric, solid organ recipients who are immunosuppressed. They were processed with the 200 ng SureSelect protocol on the Bravo automation platform and sequenced on an Illumina NextSeq. Reads were quality checked and genomes *de novo* assembled with assembly pipeline 2 (chapter 2). Table 5.4 lists the sample and sequencing information for these patients, including the samples from chapter 3. For six samples, enrichment did not work and OTR were too low to successfully assemble the genomes (samples ebv18, -20, -24, -26,- 28, -29). Consequently, there are in total 17 EBV genome sequences from immunocompromised children from the UK. Coverage plots of the succesfully assembled, additional genomes are shown in supplemental figure C.2.

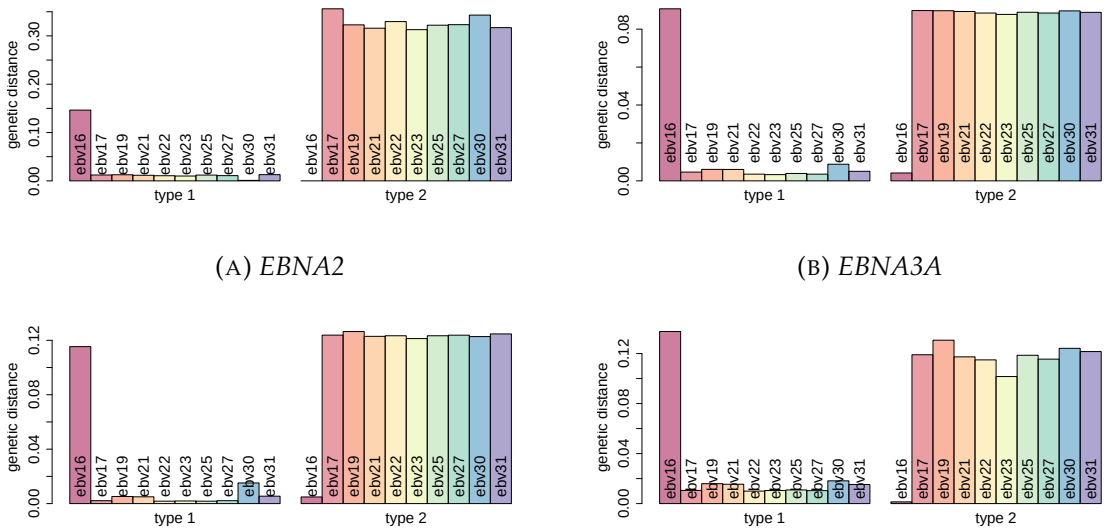
Sample	Viral load [copies/ml]	Read pairs	OTR %	Depth	Cov	
ebv6	943,000	4,150,874	1.6	38	95	●
ebv7	445,000	6,208,600	1.4	49	98	●
ebv8	1,200,000	7,277,442	1.8	84	95	●
ebv9	>2,000,000	5,509,174	9.8	450	100	●
ebv13	> 2,000,000	12,110,184	4.1	401	100	●
ebv14	> 2,000,000	18,894,696	24.9	3079	100	●
ebv15	1,295,000	11,715,202	1.4	135	99	●
ebv16	497,000	6,270,139	1.0	6	89	●
ebv17	1,067,000	6,892,091	1.9	14	100	●
ebv18	132,000	6,707,931	0.5	-	-	
ebv19	714,000	26,727,723	1.4	8	98	●
ebv20	26,000	6,477,110	0.5	-	11	
ebv21	2,525,000	5,438,799	2.5	26	100	●
ebv22	4,684,000	10,828,549	2.1	21	100	●
ebv23	1,458,000	6,203,195	1.3	7	93	●
ebv24	7,500	8,136,025	0.5	2	6	
ebv25	14,533,000	5,719,107	12.7	49	100	●
ebv26	98,000	5,790,434	0.5	2	8	
ebv27	839,000	5,220,880	2.2	34	100	●
ebv28	10,000	5,543,283	1.1	3	6	
ebv29	12,000	5,408,600	0.7	8	66	
ebv30	473,000	5,019,119	1.2	15	97	●
ebv31	1,131,000	5,460,738	1.2	10	93	●

TABLE 5.4: Sample and sequencing information for samples of paediatric, solid organ recipients under immunosuppression. OTR: On target reads in percent. Cov: Percentage of the genome that could be covered (repeats excluded). Depth: Average read depth after removal of duplicate reads. All samples that have been included for further analyses are marked by a ●.

Description of consensus sequences

In the phylogenetic tree with other type 1 EBV genomes in figure 5.1, samples do not cluster together but are spread across the part of the tree containing the other Non-Asian genomes (highlighted in orange), although some are very similar to each other.

One of the isolates (ebv16, marked with a * in figure 5.4) has a particularly long final branch, indicating a large number of changes on that branch in comparison to other sequences. Indeed, when typing the genomes, ebv16 is of type 2 given the similarity of the *EBNA2* and *-3* genes to type 1 and 2 reference strains, respectively (figure 5.3). All the other sequences belong to type 1.



(A) *EBNA2*

(B) *EBNA3A*

(C) *EBNA3B*

(D) *EBNA3C*

FIGURE 5.3: Genetic distance (K80 model) of *EBNA2* and *-3* genes of samples of immunocompromised patients from the UK to type 1 (WT) and type 2 (AG876) reference sequences.

Intrahost variation

I wanted to investigate the presence of minority variants and possibly multiple infections in the setting of immunocompromised paediatric patients. In contrast to section 5.2.1, some isolates had greater read depth due to multiple pooled sequencing runs from the optimisation experiments (chapter 3), reducing the sampling effect of medium frequency variants and allowing the detection of potential low frequency variants.

Minor variants were called with different settings depending on the depth of the respective isolate. Settings are summarised in table 5.5. A variant needed further to be supported by 2 independent reads on each strand.

Table 5.5 also lists the number of minority variants found in the samples and the fraction of the genome falling over the minimal coverage threshold. Not all samples have sufficient coverage across the genome to reliably call minority variants (e.g. ebv16, ebv17, ebv19, ebv23, ebv30, ebv31). However, there are a number of samples with sufficient

	Depth	Min cov	Min freq	frc $\geq$ Min cov	# Min var
ebv6	45	20	0.2	0.761937748	5
ebv7	57	20	0.2	0.804845535	227
ebv8	129	50	0.1	0.758456693	3
ebv9	438	80	0.05	0.83556963	0
ebv13	472	80	0.05	0.818373917	786
ebv14	3650	80	0.05	0.843352524	1
ebv15	154	80	0.05	0.724827839	71
ebv16	6	20	0.2	0.011962488	0
ebv17	14	20	0.2	0.116312643	0
ebv19	8	20	0.2	0.03275568	0
ebv21	26	20	0.2	0.646253791	0
ebv22	21	20	0.2	0.408075116	0
ebv23	7	20	0.2	0.022568646	0
ebv25	49	20	0.2	0.847747501	0
ebv27	34	20	0.2	0.715979673	0
ebv30	15	20	0.2	0.189310018	0
ebv31	10	20	0.2	0.034845477	0

TABLE 5.5: Variant call for the data set of immunocompromised patients without subsampling. Depth: average depth. Min cov: minimal depth set during variant calling. Min freq: minimal variant frequency set for variant calling. frc $\geq$ Min cov: fraction of the genome (without excluding repeats) with a depth greater or equal than the threshold. # Min var: number of high quality minority variants.

depth and coverage. In those samples, the diversity is very low with the number of minority variants ranging from $0 - 5$ (e.g. ebv6, ebv8, ebv9, ebv14, ebv21, ebv25, ebv27). There are only three samples with a high number of minority variants (ebv7, ebv13 and ebv15).

	Pos	Ref	Var	Cov	Read1	Read2	Freq	ORF	nonsyn
ebv6	143514	A	G	39	29	10	25.64		
	143552	C	T	39	11	28	28.21		
	143588	C	T	33	9	24	27.27		
	143609	G	C	26	7	19	26.92		
	143642	A	C	24	5	19	20.83		
ebv8	78258	A	C	90	54	33	36.67	BLLF1	I536S
	144203	G	A	141	111	30	21.28		
	144257	C	T	121	104	17	14.05		
ebv14	47246	A	G	3800	2166	1622	42.82		

TABLE 5.6: Minority variants of samples from immunocompromised children. Pos: WT position. Ref: Consensus base. Var: Minor variant base. Cov: Read depth at this position. Read1/2: Number of reads supporting the consensus or variant base, respectively. Freq: Minor variant frequency. ORF: Open reading frame. nonsyn: Amino acid change from consensus to variant at the respective protein position.

Table 5.6 shows the minority variants for isolates ebv6, ebv8 and ebv14, i.e. those with only few changes. The changes in ebv6 all occur in a short stretch of around 130 bp upstream of the ORF *LF3* (figure 5.4), which is associated with the BART cluster. However, they do not fall within any of the exons of the BART mRNAs nor in any of the BART

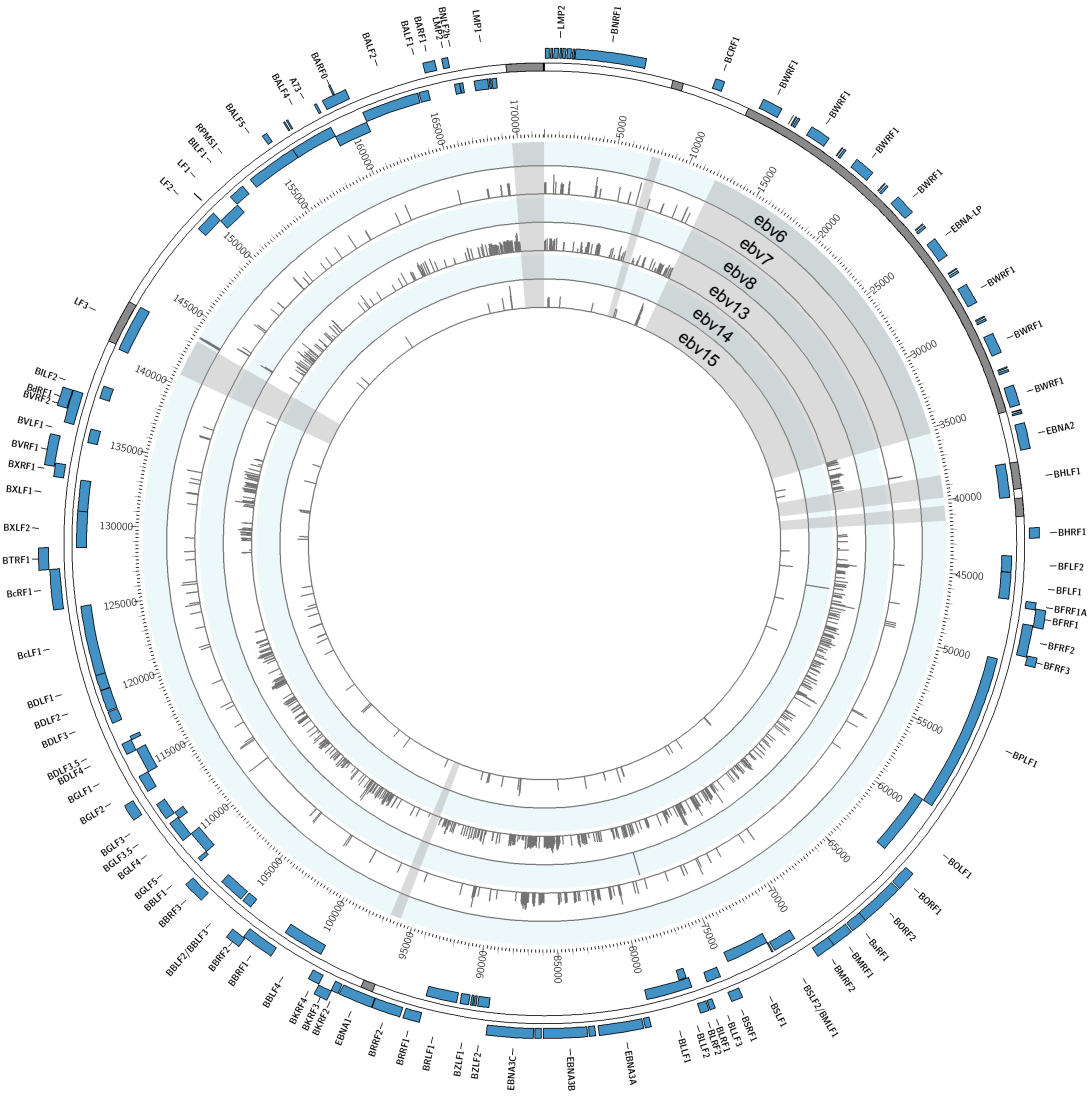


FIGURE 5.4: Genome map of all minority variants found in the data set of immunocompromised paediatric patients. Every track represents a sample, the variants are shown at their genomic position (in relation the the WT sequence) with the height of the bar indicating the variant frequency (scaled from 0 to 50 %). Repeat regions are greyed out.

miRNAs. One variant in ebv8 leads to an amino acid change (I536S) in gp350. The minority variants in these isolates likely reflect changes acquired by the infecting virus, as depth and coverage are high, but further low frequency variants especially for ebv6 and -8 cannot be excluded.

Figure 5.5 shows the variant frequency histograms of the other group of isolates with many changes. Sample ebv7 (figure 5.5a) has a bell shaped histogram, which was cut by the minimal frequency threshold of 20 %. In total, $n = 227$ minority variants were called, the majority in the frequency range of 20-30 %.

Similarly, ebv13 (figure 5.5b) has a bell shaped histogram, with the majority of variants occurring at 17-30 %. There are a few higher frequency variants between 40-50 %

and a number of lower frequency variants (< 15 %). The abundance of variants in a limited frequency range and the high number of variants ($n = 786$) suggests that these might represent a multiple infection with another EBV.

ebv7 has compared to ebv13 a much lower average depth (57 compared to 472), and it is likely more variants could be detected with deeper sequencing of the same sample.

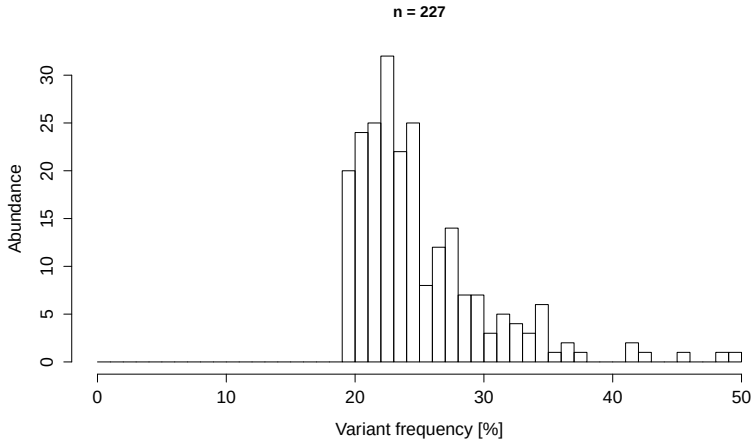
In contrast to that, the majority of variants in ebv15 (figure 5.5c) occur below 10 % and are fairly few ($n = 71$), despite high depth (154 on average), suggesting that there might either be another strain at a lower level that was not fully detected due to the limited data, or that this virus has acquired a lot of low-level variation in this patient. The relatively high number of changes compared to other samples of the same data set with very limited variation (ebv6, ebv8, ebv9, ebv14, ebv21, ebv25, and ebv27) would point towards a second low frequency strain as being the more likely explanation.

Further evidence supporting multiple infection in ebv13 and ebv7 is the presence of variants at similar frequencies spread evenly across the genome (figure 5.4).

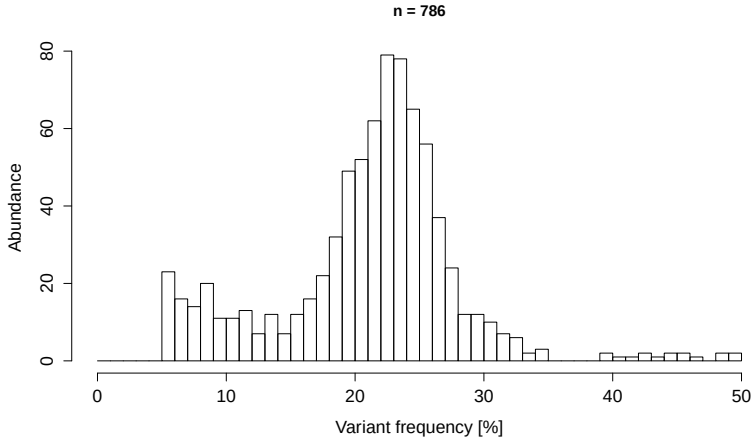
If the variant data is restricted to those with a frequency between 17 and 32 %, i.e. the main peak in figure 5.5a and 5.5b, then the minority strain of ebv13 differs at 599 positions and of ebv7 at 204 positions from the majority strain. The mean number of differences between type 1 strains is 865 ± 307 (standard deviation). At least ebv13 would therefore fall within the range of observed differences between strains; for ebv7 too much data is missing.

Of those sites, 82 and 203 are nonsynonymous in ebv7 and ebv13, respectively. The majority of minority SNPs for ebv7 affect the *EBNA3A*, *-B* and *-C* genes, both in general (figure 5.6a) as well as in the case of nonsynonymous SNPs (figure 5.7a).

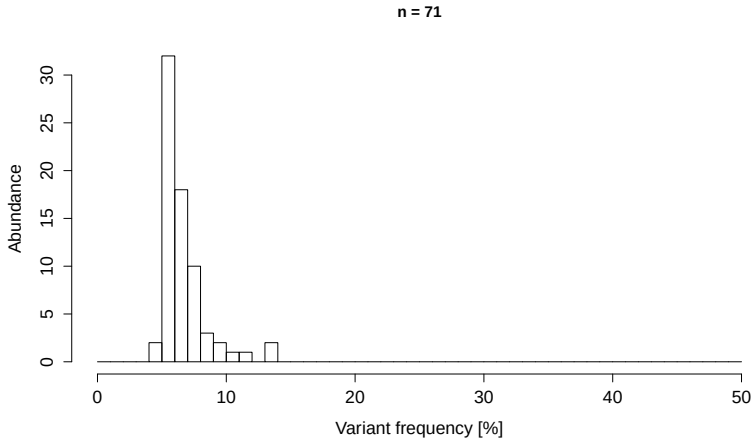
For ebv13, it is *BPLF1* (encoding for large tegument protein deneddylase), the *EBNA3*s, *LMP1*, *BLLF1* (encoding gp350) as well as *BVRF1* (encoding capsid vertex component 2) (figure 5.6b for all SNPs and figure 5.7b for nonsynonymous SNPs). Those genes belong to the most diverse genes in EBV in general, which again supports the notion that this is a genuine secondary infection with another virus.



(A) ebv7

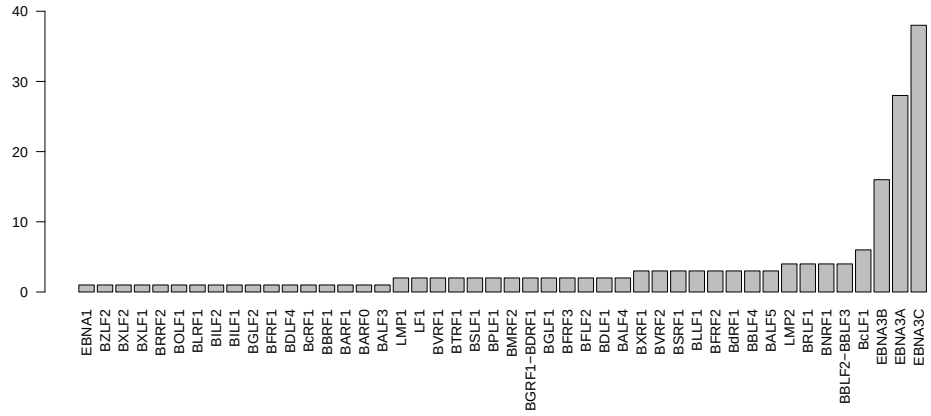


(B) ebv13

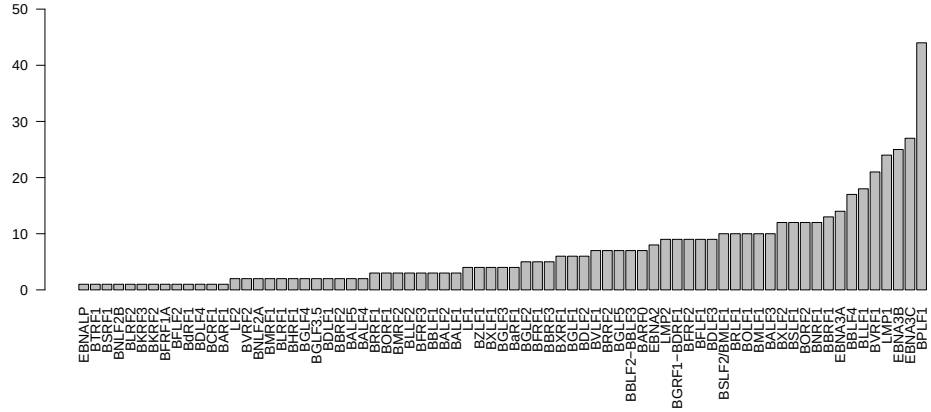


(C) ebv15

FIGURE 5.5: Variant frequency histogram of samples with greater intrahost diversity.

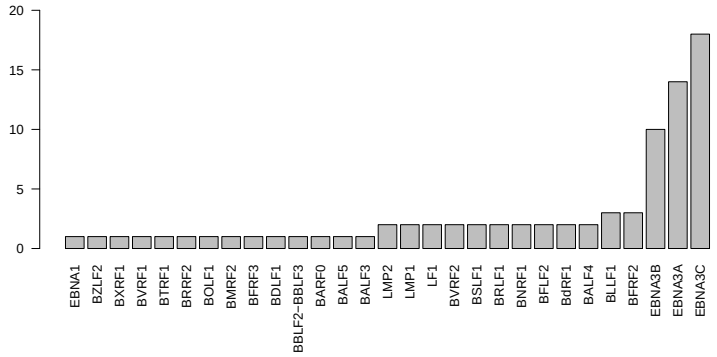


(A) ebv7, $n = 204$

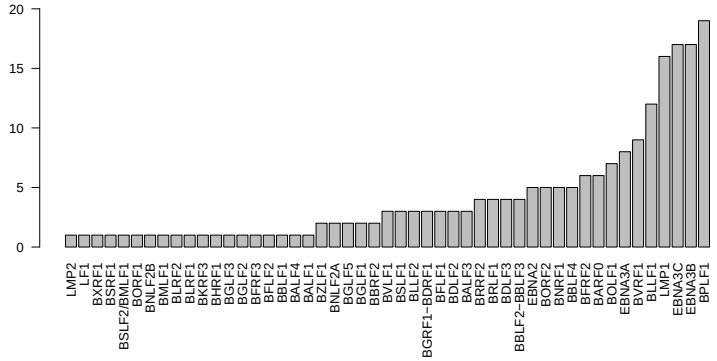


(B) ebv13, $n = 599$

FIGURE 5.6: Number of minority variants per ORF in the minority strain ($17 - 32$ %) of A) ebv7 and B) ebv13.



(A) ebv7, $n = 82$



(B) ebv13, $n = 203$

FIGURE 5.7: Number of nonsynonymous minority variants per ORF in the minority strain ($17 - 32$ %) of A) ebv7 and B) ebv13.

5.2.3 Longitudinal infectious mononucleosis samples from Japanese children

Data set

The data set comprised DNA extracts of PBMCs from 12 Japanese paediatric infectious mononucleosis patients. For every patient, two extracts were available sampled at two different time points (table 5.7). The samples were provided by Tetsushi Yoshikawa (Department of Pediatrics, Fujita Health University School of Medicine, Toyoake, Japan) and are a subset of the data presented in Nakai et al., 2012.

Patient	Sample	Age	Days after onset	Viral load [copies/µg]	Read pairs	OTR %	Depth	Cov	
1	P1-812	6	6	114,000	3,803,104	3.2	25	98	●
	P1-833		17	240	-	-	-	-	
2	P2-1213	7	7	125,000	20,113,834	0.4	37	94	●
	P2-1246		19	14,000	5,368,060	0.9	8	85	●
3	P3-2670	7	5	250,000	3,672,286	4.1	81	96	●
	P3-2740		28	17,000	7,536,334	1.0	13	92	●
4	P4-2274	6	10	139,000	3,435,030	8.1	16	100	●
	P4-2392		59	4,000	5,972,994	1.6	19	97	●
5	P5-1294	15	8	60,000	3,079,348	2.6	62	100	●
	P5-1323		15	7,000	1,771,093	0.7		29	
6	P6-1751	3	4	50,000	3,832,812	3.8	31	96	●
	P6-1789		16	350	1,811,779	0.1	-	-	
7	P7-2315	11	12	10,000	3,235,942	0.6	7	32	
	P7-2634		84	24,000	1,138,767	1.8	6	80	●
8	P8-414	7	3	22,000	3,628,334	7.8	76	100	●
	P8-516		66	9,000	6,724,909	2.0	23	95	●
9	P9-2631	4	9	63,000	3,262,866	5.5	66	100	●
	P9-2645		14	38,000	6,875,569	5.2	97	100	●
10	P10-2187	1	10	4,000	3,425,692	0.1	8	27	
	P10-2777		211	210	1,299,344	0.1	2	6	
11	P11-871	1	7	24,000	1,703,065	14.2	22	100	●
	P11-920		30	3,500	-	-	-	-	
12	P12-1026	11	15	3,000	2,857,532	0.9	32	90	●
	P12-1078		42	1,000	1,743,347	0.1	-	-	

TABLE 5.7: Sample and sequencing information for longitudinal samples of infectious mononucleosis samples from Japanese children. OTR: On target reads in percent. Cov: Percentage of the genome that could be covered. Depth: Average read depth after removal of duplicate reads. All samples that have been included for further analyses are marked by a ●.

All samples were processed with the 200 ng SureSelect protocol on the Bravo automation system with a 1:10 dilution of capture baits (0.2 µl) and were sequenced on a MiSeq in two batches. As the OTR output was still fairly poor for many samples, the library preparation and hybridisation was repeated on the Bravo automation system and sequenced on a NextSeq in one run to achieve higher read depth. A few samples were excluded if there was not enough input material left or if viral loads were so low that the enrichment success was unlikely (P1-833, P6-1789, P10-2277, P11-920 and P12-1078). Additionally, the sequence reads from the optimisation experiment in chapter 3 for samples P2-1213 and P2-1246 were included (manual preparation with the 3 µg and 200 ng protocol with varying bait dilutions). A summary of the sample preparation is given in table C.1.

Read data for the same samples were pooled after QC and *de novo* assembled using the assembly pipeline 2 (see chapter 2).

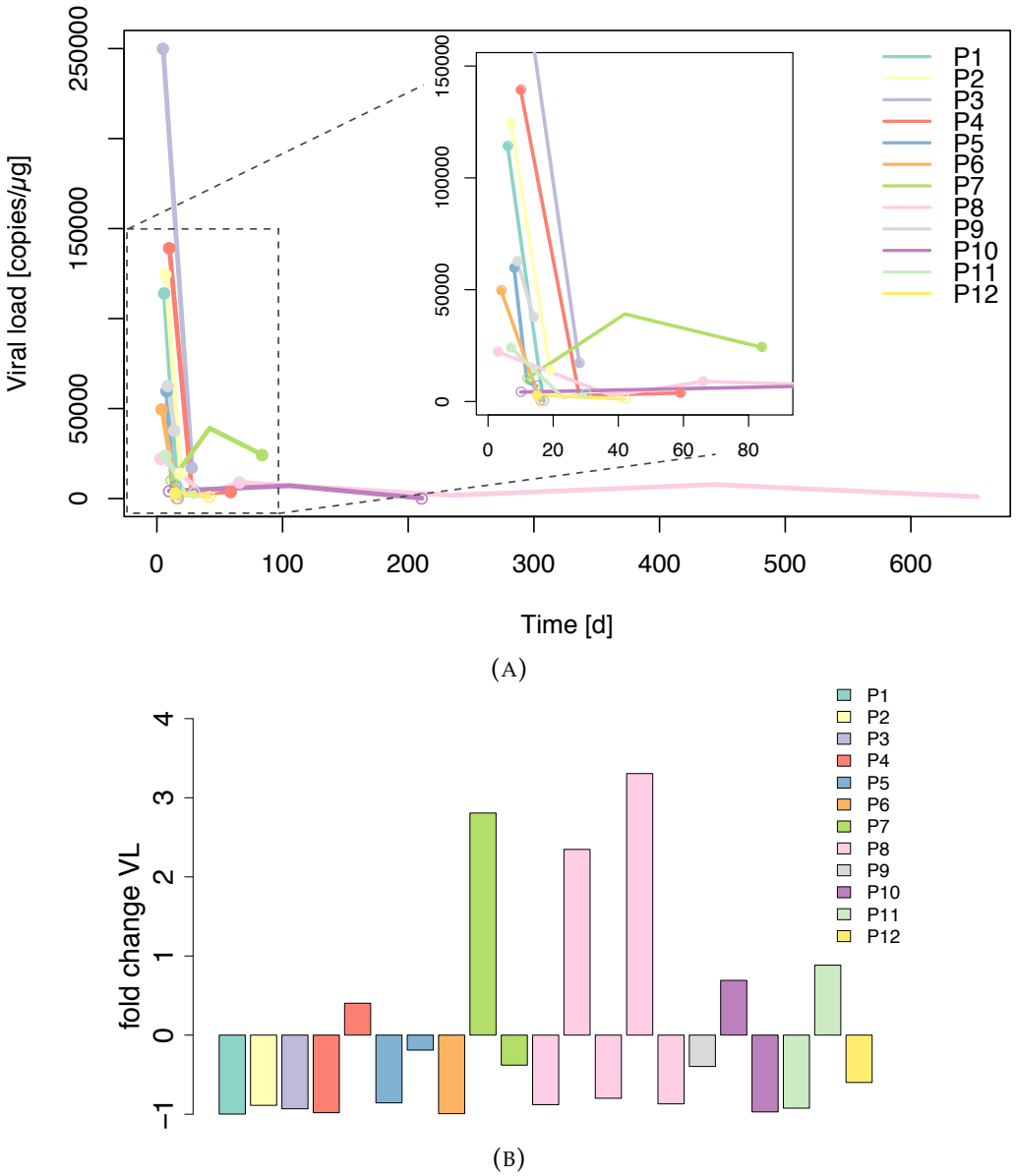


FIGURE 5.8: A) Viral load of all twelve IM patients given in copies/µg DNA. Time points, for which DNA samples are available, are marked with an open circle; samples for which genomes could be assembled are marked with a filled circle; time points without any specific symbol only have a VL measurement, but no DNA sample was available. B) Fold change of viral load from one time point to the next.

Table 5.7 shows the assembly results and coverage plots are shown in supplemental figure C.3. For some samples, the second time point did not yield any results as viral titres were too low (figure 5.8a) and consequently also the number of OTRs (table 5.7, P5-1323 and P7-2315). In total, 16 samples were retained, with longitudinal, paired samples for five patients.

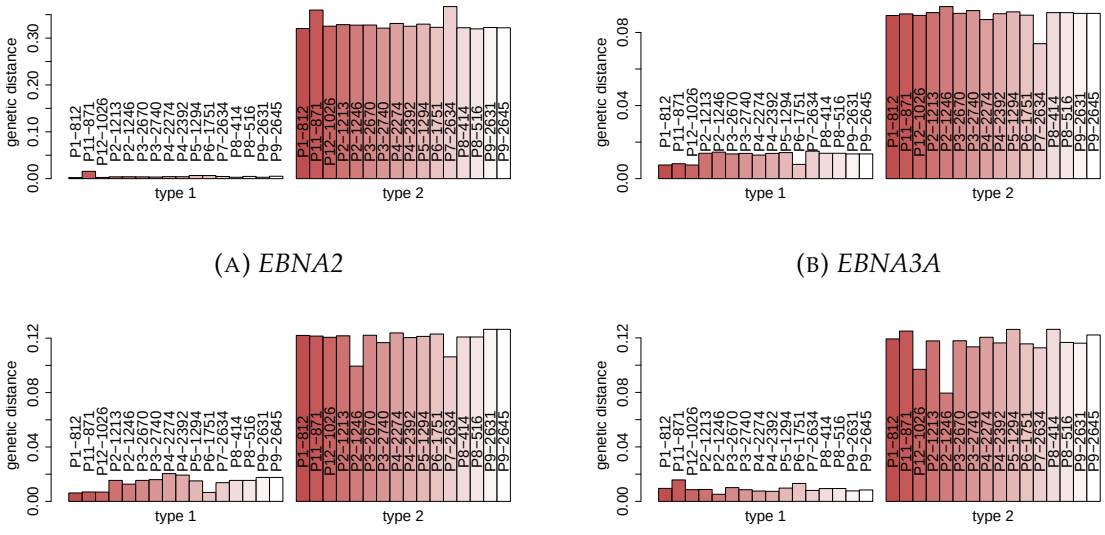
Figure 5.8 depicts the viral load (VL) of patients across time, including time points

for which there was a VL measurement but no DNA sample. The VL drops for many patients fairly quickly, i.e. the host controlled the EBV infection successfully. This can be seen more clearly in figure 5.8b by means of the fold change from time point to time point. A fold change of -1 indicates a nearly complete control of EBV, such as in P1 to P6. Some patients, however, retain elevated VL (especially P7, but also P8) suggesting they do not control their EBV infection. But the trajectory of VL for the other samples is not always clear due to variable number of data points for the VL measurements.

In Nakai et al., 2012, patients have been divided into two groups: slow and fast regression. The former was defined as having >500 copies/µg DNA 21 days after onset of illness. According to this classification, P1 and P6 belong to the fast regression group and all other patients belong to the slow regression group. In the same publication, no statistically significant differences between the groups were found in clinical features, but during acute phase, the fast regression group showed higher serum concentrations of interleukin (IL)-1β, IL-12, tumour necrosis factor-α, interferon-inducible protein 10 and monokine induced by interferon γ (Nakai et al., 2012).

Description of consensus sequences

All patients are infected with EBV type 1 as all samples have smaller distances to type 1 than to type 2 typing genes of the respective reference strains WT and AG876 (figure 5.9).



(A) *EBNA2*

(B) *EBNA3A*

(C) *EBNA3B*

(D) *EBNA3C*

FIGURE 5.9: Genetic distance (K80 model) of *EBNA2* and *-3* genes of infectious mononucleosis samples to type 1 (WT) and type 2 (AG876) reference sequences.

In a phylogenetic tree with additional type 1 genomes, all samples fall within the previously determined Asian cluster (figure 5.1, highlighted in blue, IM samples coloured in red). They also cluster together with the other Asian samples in a split network (supplemental figure C.4) and are assigned to the same blue, Asian subpopulation as the other

Asian derived samples (supplemental figure C.5). Sequences of the same patients usually cluster together in the tree. The only exceptions are the two samples of P2. They appear together in a well-supported group of samples that are very similar to each other together with P3, P8 and a gastric carcinoma-cell line from South Korea (YCCEL1/LN827561).

Short branch length between samples of the same patient indicate only few variations over time. Table 5.8 lists the consensus level changes found in three patients.

P2 ($\Delta t = 12$ d) had four variations, three of them close together in the ORF *LF2* that affect together two amino acid changes (I77S and L78V).

P3 ($\Delta t = 23$ d) and P9 ($\Delta t = 5$ d) only showed one SNP difference. The one in P3 lies in the overlapping *BARF0* and *BALF3* ORFs and leads to an amino acid change in both gene products (L304P for BARF0, a putative protein, and Q397R for BALF3, the tripartite terminase subunit 1, respectively), while the one in P9 is not in a coding region. All of these samples belong to the slow regression group, but P2 and P3 seem to control their infection (figure 5.8), while the trajectory of P9 is less clear, but the VL is dropping by 0.39-fold within five days.

	140489	150108	150109	150110
P2-1213	C	C	A	C
P2-1246	A	G	T	A
nonsyn		•	•	•

(A) Patient 2.

	160046
P3-2670	C
P3-2740	T
nonsyn	•

(B) Patient 3.

	140989
P9-2631	C
P9-2645	A
nonsyn	

(C) Patient 9.

TABLE 5.8: SNPs in longitudinal data of IM on consensus level. The isolates are ordered by sample date. Ambiguity codes are R= G/A, S=C/G. N denotes missing data. Positions are patient WT coordinates.

Description of intrahost variation

In order to study the intrahost variation in primary infection of immunocompetent children, variants were called as described with cut-off values for minimal depth set to 20 and minimal frequency to 0.2. Not all samples had sufficient depth to allow variant calling across the majority of the whole genome (table 5.9). The number of variants ranged between samples from 0 to 60. Some samples (P1, P4 and especially P12) seem to be fairly diverse even with only a smaller fraction of the genome having sufficient depth. It is likely a lot of the true diversity is being missed in these samples due to missing data. All minority variants are shown in the EBV genome map in figure 5.10. The complete list of variants is given in supplemental table C.2, including the affected ORFs and nonsynonymous changes.

P12-1026 harbours the most minority variants and they mostly fall within the generally more diverse genes, such as the *EBNA3*s, *LMP2*, *BRRF2* (encoding a tegument protein) and *BKRF2* (encoding the envelope glycoprotein L). The distribution of variant frequencies is shown in figure 5.11 and resembles a bell shape (cut off at the lower side

	Depth	frc ≥ min cov	# Min var
P1-812	25	0.524905843	15
P2-1213	37	0.769889456	6
P4-2274	16	0.241898397	31
P5-1294	62	0.800502948	1
P6-1751	31	0.71997299	21
P8-516	23	0.483983072	0
P9-2631	66	0.841472288	1
P11-871	22	0.505614511	1
P12-1026	32	0.557283147	60

TABLE 5.9: Variant call for the data set of IM patients. Depth: average depth. frc ≥ Min cov: fraction of the genome (without excluding repeats) with a depth greater or equal than the threshold. # Min var: number of high quality minority variants.

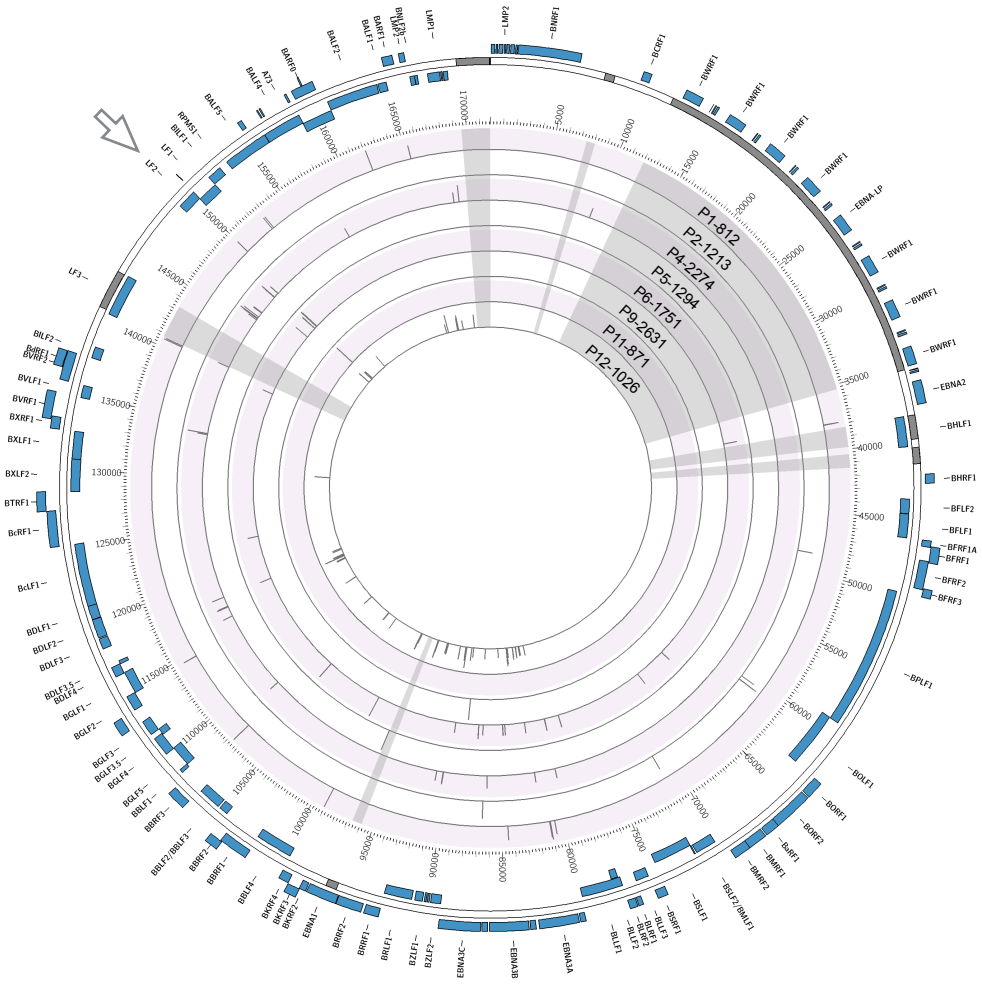


FIGURE 5.10: Genome map of all minority variants found in the data set of IM patients. Every track represents a sample, the variants are shown at their genomic position (in relation the the WT sequence) with the height of the bar indicating the variant frequency (scaled from 0 to 50%). Repeat regions are greyed out.

due to the 20 % cut-off threshold). Only around 55 % of the genome had enough data to call variants with the chosen thresholds. Of the 60 variant sites, 20 were nonsynonymous (table C.2). It is hard to asses whether this is truly a mixed infection. The relatively low number of variant sites ($n = 60$) and the fact that they mostly cluster in genetically diverse areas indicate it could rather be due to infection with an already relatively similar strain and/or intrahost evolution.

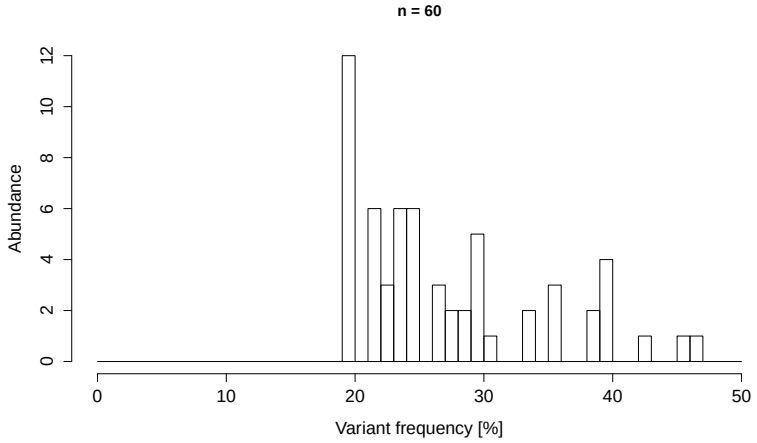


FIGURE 5.11: Variant frequency histogram of P12-1026.

Interestingly, in five of the eight genomes for which minority variants could be detected, *LF2* was affected at in total five loci (marked with an arrow in figure 5.10, and specifically labelled in table C.2). Two sites (149146 and 149326) were each shared by two isolates (P4/P12 and P6/P12, respectively). On the consensus level, four of the five alleles do also occur in other published genomes, whereas one was a singleton. This is the same ORF for which two longitudinal nonsynonymous consensus level changes in P2 were detected (but these exact loci were not detected to be heterogeneous at these frequency/quality cut-offs).

5.2.4 Comparison of intrahost diversity between immunocompetent and immunocompromised paediatric patients

While there are six other IM-derived EBV genomes published (the WT strain (NC_007605) (Baer et al., 1984; de Jesus, 2003) and LN827596, LN827590, LN827567, LN827799 and LN827583 (Palser et al., 2015)), they are all from LCLs. The IM data set here represents the first genome set of primary infection derived directly from blood. As such, it allows the comparison of intrahost diversity of clinical sample-derived sequences between immunocompetent (IM) and immunocompromised paediatric (paired tumour and blood sample (TB) from section 5.2.1 and other immunocompromised (IC) from 5.2.2) patients.

Figure 5.12 shows the intrahost diversity (π_i) of the different data sets. The value represents an average across each sample (see chapter 2). The TB set was split into the tumour (TB-T) and blood-derived (TB-B) sequences. The tumour data refers to the subsampled data (as described in section 5.2.1). Supplemental figure C.6 shows the same

data including the complete data set for the tumours (TB-Tall); here, diversity scores were much higher ($\overline{\pi}_{i,TB-Tall} = 0.001$ vs. $\overline{\pi}_{i,TB-T} = 0.00013$). This is probably due to not excluding the repeat regions in the calculations, which are based on the whole mapping file (reads against consensus). While the diversity is calculated under consideration of sequencing and PCR errors as well as mapping quality filtering, repeat regions likely accumulate lots of high quality mapping reads, whose position (which repeat unit) is not reliable. Moreover, no minority variants were detected (repeat regions were excluded) in either the subsampled or complete data (with one exception) (see table 5.3). To correct for this and also to ensure comparability of data in terms of depth, the subsampled data set was chosen.

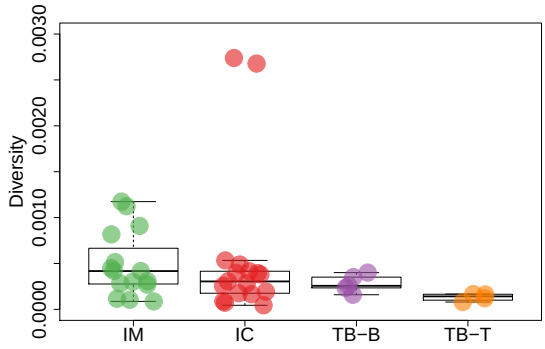


FIGURE 5.12: Intrahost diversity across different data sets. The line in the box indicates the median, the bottom and top of the box show the 25th and 75th percentile. IM: Infectious mononucleosis; IC: immunocompromised patients, TB: paired tumour (-T) and blood (-B) data set. The TB-T data set represents the subsampled data set of comparable depth to its paired blood samples.

Diversity was higher in the IM data set ($\overline{\pi}_{i,IM} = 0.00048$) compared to the paired tumour and blood samples (TB-B and TB-T, $\overline{\pi}_{i,TB-B} = 0.00028$ and $\overline{\pi}_{i,TB-T} = 0.00013$), but not the immunocompromised (IC) ($\overline{\pi}_{i,IC} = 0.00056$) samples. This is due to the two outliers in IC, as the distribution of the rest of the samples is generally lower. Accordingly, the median of IC (0.0003) is lower than that of IM (0.00042). Differences were not significant (IM vs IC: $p = 0.35$, IM vs TB-B: $p = 0.23$, Mann-Whitney U test) except between IM and tumour samples (IM vs TB-T: $p = 0.037$). There was also no significant difference between IM and IC even when excluding the two outliers in IC (named hereafter IC*), which correspond to the multiply infected samples ebv7 and ebv13 ($\overline{\pi}_{i,IC*} = 0.00028$, $p = 0.097$).

Blood-derived samples (IC and TB-B) were similar ($\overline{\pi}_{i,TB-B} = 0.00028$ and $\overline{\pi}_{i,IC*} = 0.00028$), as expected, as both groups are representative of immunocompromised patients. But the diversity in tumours was much lower ($\overline{\pi}_{i,TB-T} = 0.00013$), and although differences were not significant, p-values are indicative of a potential difference between groups (IC vs TB-T: $p = 0.05$, TB-B vs TB-T: $p = 0.063$, Mann-Whitney U test).

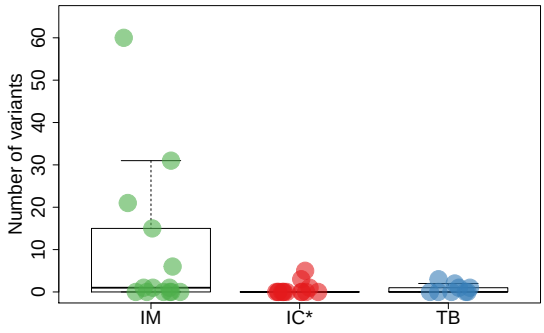


FIGURE 5.13: Number of minority variants for different data sets. IM: Infectious mononucleosis, IC*: immunocompromised patients without multiply infected samples (ebv7, ebv13, ebv15), TB: paired tumour and blood samples (not subsampled).

This is supported by the number of minority variants in the different data sets (figure 5.13, multiple infected samples in IC excluded), although differences between groups are not significant ($p = 0.1123$, Kruskal-Wallis rank sum test). However, it is not clear within the IM samples, whether the higher diversity results from multiple infection or within host evolution, as many of the samples with higher number of variants have poor depth or coverage across the genome.

5.2.5 Comparison between Asian NPC and Non-NPC genomes

The majority of published genome sequences from Asia are isolates from NPC samples ($n = 14$). There are six genomes from Non-NPC settings: three sLCL from Hong Kong, one BL from Japan, one GC from South Korea and one BL from Papua New Guinea (see table 1.2). The genome from Papua New Guinea is the only type 2 genome in this group. The IM samples sequenced in this project are therefore the first non-tumour EBV isolates from an Asian region where NPC is not endemic (Kimura et al., 2011).

Determining whether there are differences between Asian NPC and other Asian Non-NPC samples might help to disentangle geographic from malignancy-associated variation. The type 2 genome has been excluded as the *EBNA* genes would bias the results. The lower quality sequence of longitudinal pairs from the Japanese IM set have also been excluded. A split network of these $n = 30$ Asian NPC ($n = 14$) and Non-NPC ($n = 16$) genomes (figure 5.14) shows that the majority of NPC-derived genomes cluster together. The reticulations, however, also indicate that recombination between strains has likely occurred. In order to determine relevant SNPs responsible for this (partial) segregation, a PCA was performed. Figure 5.15a shows the scatterplot of the PCA and figure 5.15b a NJ tree of the same genomes as a complementary identification key for the scatter plot.

The first principal component (PC1) explains around 40 % of the variation in the data. There is a rough separation between NPC and Non-NPC samples with two clusters on the far left (only NPC samples) and far right (mostly Japanese IM samples, the South Korean GC samples (LN827561) as well as GD1, a saliva sample from a Chinese NPC

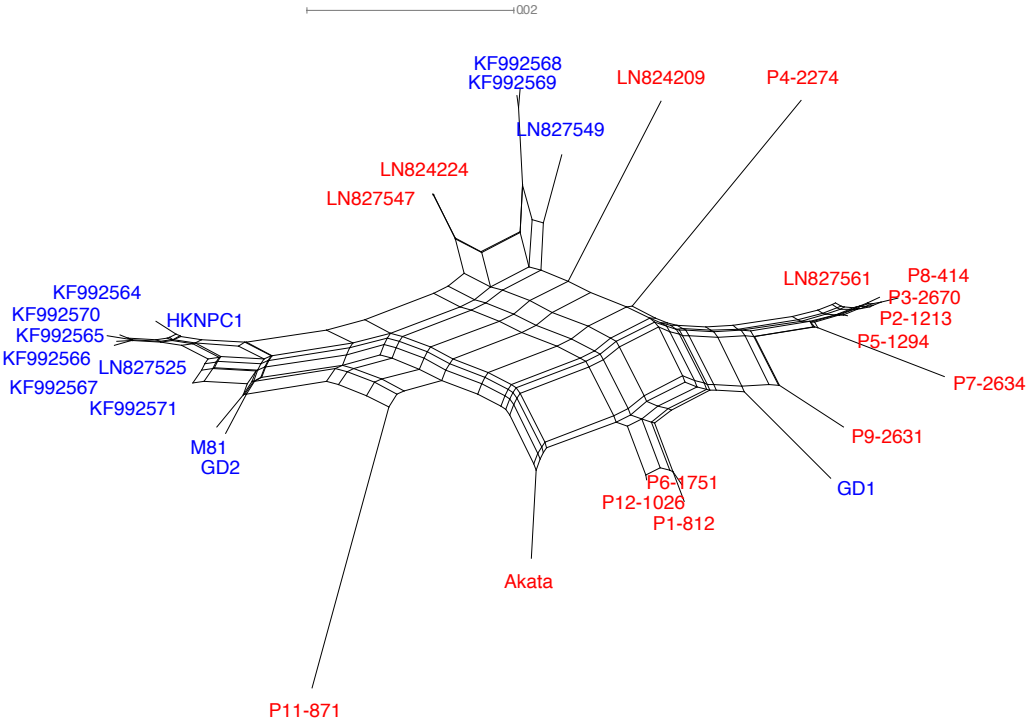


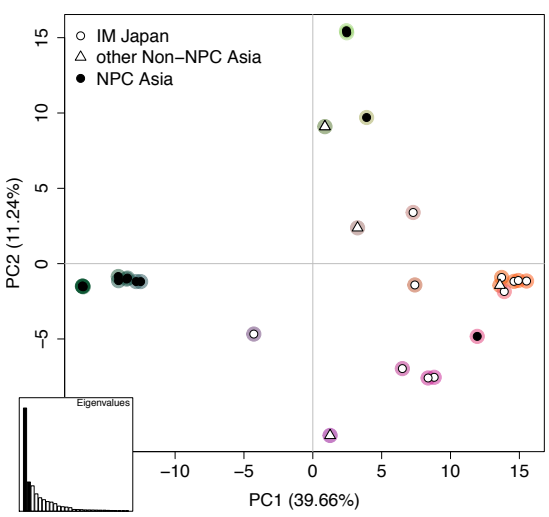
FIGURE 5.14: Split network of Asian NPC (blue) and Non-NPC (red) derived genome sequences.

patient). The second principal component (PC2) only explains around 11 % of the variance in the data, but seems to explain the variation within the intermediate samples of PC1. These intermediate samples comprise the three sLCL from Hong Kong (LN827547, LN824224, LN824209), the BL from Japan (Akata) as well as some IM samples from Japan and Southern Chinese NPC samples.

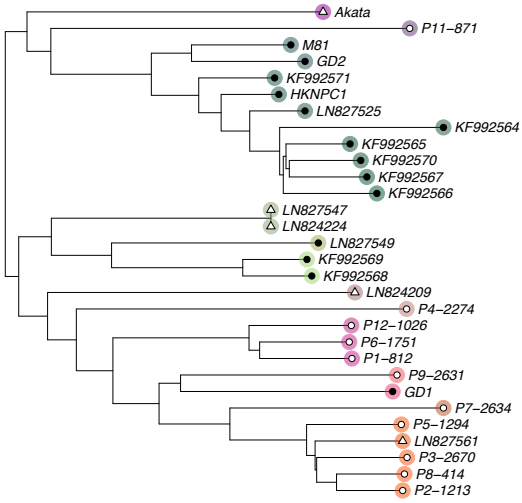
The loadings of the variables from the PCA allow us to determine the variables (SNPs) with the greatest contribution to PC1 (i.e. these SNPs are responsible for the divide between the two clusters on axis 1). Figure 5.16 shows the SNP locations in the genome whose absolute loadings are higher than the third quartile of all absolute loadings. There are 767 SNPs that fulfil this criterion (red track), the majority of those, 626, lie within coding regions (blue track). Of these, 271 SNPs lead to a change of amino acid in the encoded proteins (green track). Figure 5.17 shows the number of nonsynonymous substitutions for each affected ORF. The largest number of nonsynonymous SNPs lie within the ORFs *BPFL1* and the *EBNA3*s, followed by the inner tegument protein gene *BOLF1*, the gp350 gene *BLLF1* as well as *LMP2* and *BZLF1*.

Except for *BPLF1*, all of these ORFs encode antigenic proteins. In fact, 24 of the nonsynonymous changes lie within experimentally described epitopes (*Immune Epitope Database*): six in *BZLF1*, six in *EBNA3A*, five in *LMP2*, four in *EBNA3C*, and three in *EBNA3B* (supplementary table C.3).

In track A of figure 5.16, track A, between 138 kb to 148 kb, there are a number of SNPs in the noncoding area between *BILF2* and *LF3* as well as between *LF3* and *LF2*,



(A) PCA scatterplot of the first two principal components (PC), accounting for ≈ 40 % and 11 % variance, respectively, in the data.



(B) Midpoint-rooted NJ tree coloured with by the respective PCA scores as a identification key for the PCA scatterplot.

FIGURE 5.15: PCA of whole genomes of Asian NPC and Non-NPC samples.

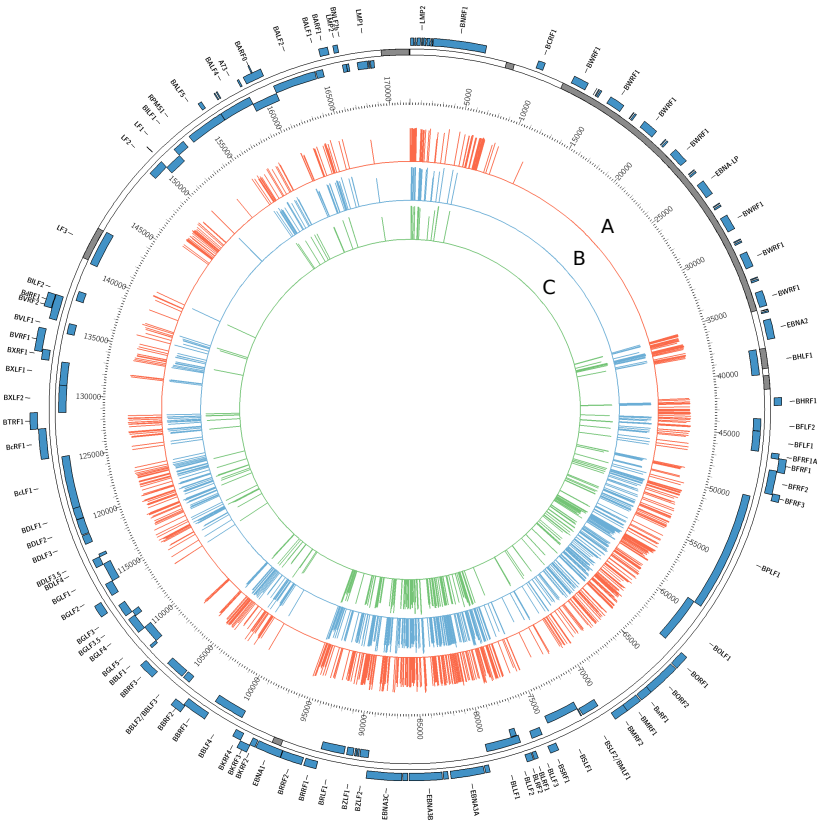


FIGURE 5.16: Absolute loadings of the variables that are greater than the third quartile plotted around the genome map. A: all SNPs, $n = 767$; B: all SNPs within coding regions, $n = 626$; C: nonsynonymous SNPs, $n = 271$.

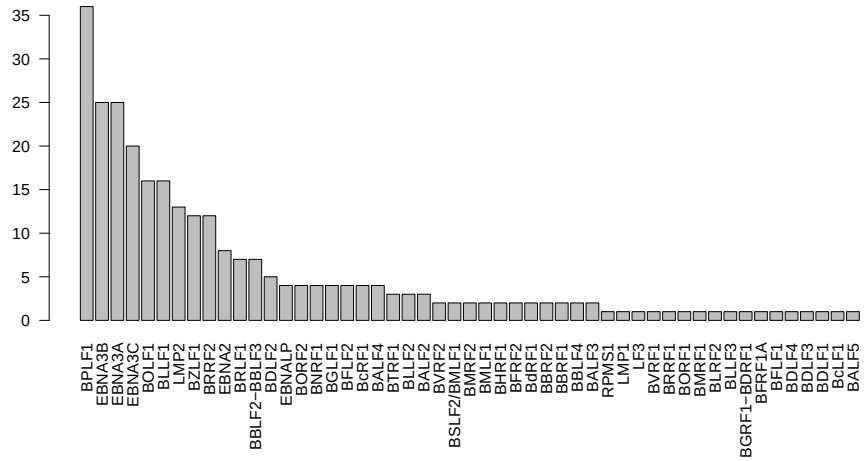


FIGURE 5.17: Number of nonsynonymous substitutions per gene in the group SNPs that have the highest impact on variation explained by PC1 in Asian NPC and Non-NPC genomes.

corresponding to the location of the BamHI A rightward fragment (BART) from which mRNAs and miRNAs are transcribed and which are strongly expressed in NPC and thought to play a role in pathogenesis (Edwards, Marquitz, and Raab-Traub, 2008; Cai et al., 2006).

None of the SNPs of interest lie within the miRNAs, which are very conserved. Ten SNPs however fall into the exons of the BARTs (exon coordinates were given in Yamamoto and Iwatsuki, 2012). These exons are V, VA' and VB, which are also overlapping with the ORFs *BALF4* and *BALF5*. In addition, there is a single SNP in the first exon of *RPMS1*, a putative ORF also part of the BART family transcripts.

The next question was then if any of the nonsynonymous sites are unique to the cluster of NPC samples (figure 5.15, samples M81, GD2, KF992571, HKNPC1, LN827525, KF992564, KF992565, KF992570, KF992567, KF992566) when the remaining Non-Asian type 1 sequences are included. For this, the consensus sequence of this cluster was taken based on the 271 nonsynonymous sites and compared to all other sequences. Only eight of these sites were unique to the NPC cluster: 976 and 2086 in *BNRF1*, 60 and 65 in *EBNALP*, 1828 in *BPLF1*, 2087 in *EBNA3B*, 1279 in *BRRF2*, and 375 in *BGLF1* (CDS positions). The majority of sites do not lie within any known or in the UniProt database annotated structural/functional areas. Only the SNP in *EBNA3B* lies within a proline rich region. The variant found in the NPC cluster is on the DNA level a T instead of C, and on the amino acid level isoleucine instead of threonine/alanine (both variants occur in the other sequences due to additional variations at different codon positions).

Other NPC-associated variations I also looked for other variations that have been described in the literature to be potentially associated with NPC.

BZLF1 major switch from latency to lytic. Difference in epitheliotropism might be related to different expression levels of this gene. Many regulatory elements in the *BZLF1* promotor (Zp) have been found in a short stretch ranging from -221 to +12 relative the

transcription start (WT coordinates 91,126 to 90,893). Four groups of Zp variants have been described: Zp-P, which is identical to the WT sequence, as well as Zp-V1, Zp-V3 and Zp-V4 based on four positions (-196, -141, -106 and -100) (Lorenzetti et al., 2009).

In the data set here of Asian type 1 sequences, 18/30 were of type Zp-V3 (TGGG) and 12/13 were of type Zp-P (TAAT). Zp-V3 was more often observed in NPC cases (11/14) than in non-NPC (7/16) cases. Conversely, Zp-P was found more often in non-NPC (9/16) than NPC (3/14) cases. The null hypothesis that NPC pathology and promotor variant are independent cannot be rejected, however ($p = 0.052$, Chi-square test).

Another arguably NPC associated variation is a 30 bp deletion in *LMP1*. All genomes except patient 4 of the IM data set exhibited this. Moreover, the XhoI restriction site seems to be present in all genomes (where data available). Interestingly, there is only one SNP in the *LMP1* ORF, that belongs to the group of sites contributing most to the partitioning of the two PCA clusters (figure 5.16).

		intergenic										EBER2							
	group	6809	6855	6857	6867	6885	6887	6912	6921	6935	6945	6982	7000	7013	7017	7024	7049	7124	variant
WT		T	G	G	G	G	T	A	C	G	G	C	T	A	A	G	A	A	B95-8
Akata	Non-NPC	A				A	G	G		C	A							G	
GD1	NPC	A				A	G	G			A							G	
LN824209	Non-NPC	A	A			A	G	G	A		A							G	
LN827561	Non-NPC	A				A	G	G			A	A						G	
P1-812	Non-NPC	A				A	G	G			A							G	
P11-871	Non-NPC	A				A	G	G			A					A		G	
P12-1026	Non-NPC	A				A	G	N	N		A							G	
P2-1213	Non-NPC	A				A	G	G			A							G	EB-6m
P3-2670	Non-NPC	A				A	G	G			A							G	
P4-2274	Non-NPC	A				A	G	G			A							G	
P5-1294	Non-NPC	A				A	G	G			A							G	
P6-1751	Non-NPC	A				A	G	G			A							G	
P7-2634	Non-NPC	A	N	N	N	A	G	G			A							G	
P8-414	Non-NPC	A				A	G	G			A							G	
P9-2631	Non-NPC	A				A	G	G			A							G	
GD2	NPC	A			A								G	-	-	-	C	G	
HKNPC1	NPC	A		T	A								G	G	T		C	G	
M81	NPC	A			A								G	G	T		C	G	
KF992564	NPC	A			A								G	G	T		C	G	
KF992565	NPC	A			A								G	G	T		C	G	
KF992566	NPC	A			A								G	G	T		C	G	
KF992567	NPC	A			A								G	G	T		C	G	
KF992568	NPC	A			A								G	G	T		C	G	EB-8m
KF992569	NPC	A			A								G	G	T		C	G	
KF992570	NPC	A			A								G	G	T		C	G	
KF992571	NPC	A			A								G	G	T		C	G	
LN824224	Non-NPC	A			A								G	G	T		C	G	
LN827525	NPC	A			A								G	G	T		C	G	
LN827547	Non-NPC	A			A								G	G	T		C	G	
LN827549	NPC	A			A								G	G	T		C	G	

FIGURE 5.18: Variations found in the EBER region (genes and intergenic region). Positions and changes are indicated in relation to WT. N denotes missing data.

The EBV encoded small RNAs (EBER) *EBER1* (167 bp) and *EBER2* (173 bp) are the most abundant viral transcripts (Rymo, 1979). They are expressed in all EBV-associated tumours. Both EBER genes are highly conserved among the EBV lineages. The majority

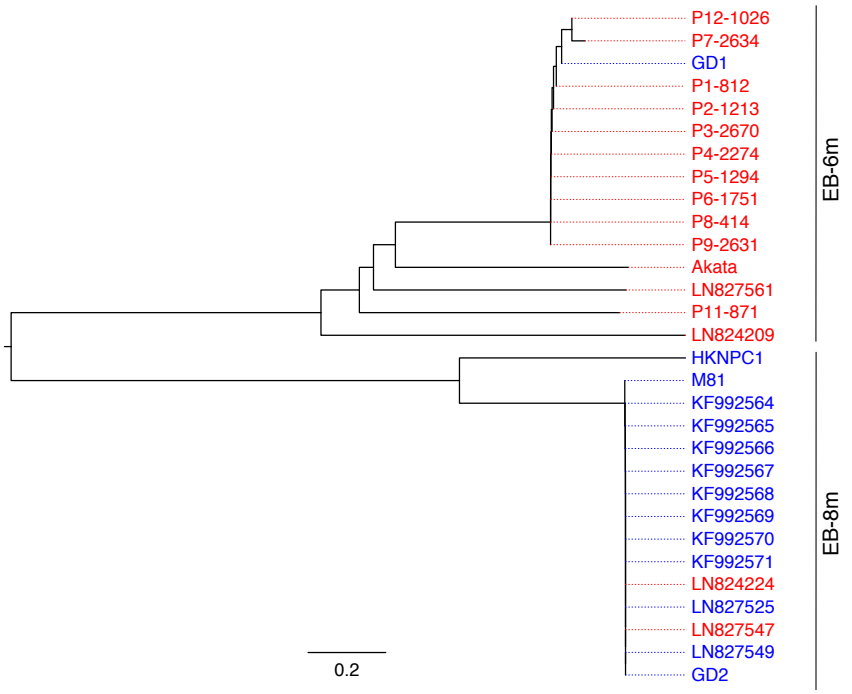


FIGURE 5.19: Midpoint-rooted NJ tree of the EBER region including intergenic region of Asian NPC (blue) and Non-NPC (red) samples.

of the variation can be found in the 161 kb spacer region. In the data set here, *EBER1* shows no variation. Within the whole EBER region including the spacer, there are 17 polymorphic sites; seven of them lie within *EBER2*, 10 of them in the spacer region.

Figure 5.18 summarises the SNPs found between the Asian NPC and Non-NPC samples. A ML tree had too little support due to the lack of variation (except for the two most basal nodes), but the NJ tree of the whole EBER region (including spacer) in figure 5.19 recapitulates this pattern suitably as well. There are four defined variant types of EBERs, B95-8 as found in the WT sequence, EB-8m, EB-10m and EB-6m (Wang et al., 2010c). 15 isolates are of type EB-8m and 15 isolates are of type EB-6m (figure 5.18). Type EB-8m was found most frequently in NPC (13/14) than in Non-NPC samples (2/16), whereas EB-6m was found most frequently in Non-NPC (14/16) than NPC samples (1/14). This difference is significant ($p = 1.13\text{e}{-5}$, Chi-square test).

5.3 Discussion

5.3.1 EBV infection in paediatric immunocompromised patients

Here, I sequenced for the first time paired tumour and longitudinal blood samples from patients, in this case paediatric PTLD cases. Additionally, a number of EBV genomes from blood of immunocompromised paediatric patients were sequenced. Whole genome sequencing of EBV directly from blood is still a greater challenge than from tumours due to the great difference in viral load. As a consequence, the greatest drawback of the analyses presented here is the limitation due to the reduced depth achievable from blood

samples. In the case of immunocompromised patients, this is further hampered by the administration of Rituximab, a monoclonal antibody against the B cell surface antigen CD20 that results in B cell depletion, the main site of EBV infection. Rituximab is used both in prophylaxis and treatment for PTLD after HSCT and SOT (Heslop, 2009; van der Velden et al., 2013; Blaes et al., 2005; Oertel et al., 2005).

Sequences from paired tumour and blood sample of the same time point from the same patient resemble each other strongly, suggesting that the majority strain found in the blood is established in the tumour. This is concordant with the hypothesis that the progenitor cells of the tumour have already been infected prior to malignant transformation (Capello, Rossi, and Gaidano, 2005). Moreover, the virus population in the tumour is extremely homogeneous.

For the two cases where longitudinal data was available, the majority strain in the blood remains the dominant virus in the blood for many years. While the sample size is rather small, the results are still indicative that variation of EBV over time without multiple infection in immunocompromised patients at the consensus level might be rather limited. However, CTL escape mutants after donor CTL infusion in a case of PTLD have been described before in one patient (Gottschalk, 2001), but selective pressure might be low due to EBV's ability to establish latency in the advent of immune control.

At the minority level, in both data sets (TB and IC), no variation was found in the majority of samples, though in some cases this might be attributed to insufficient depth and coverage. Even samples with fair or high depth, however, variable sites were usually few. Those were mostly synonymous changes; the few nonsynonymous changes were interestingly mostly observed in genes that encode for structural proteins of the capsid (*BGLF1*, *BORF1*) and glycoproteins (*BLLF1*, *BILF2*). These could potentially be related to immune escape, but in immunocompromised patients, latent replication is usually the origin of high viral loads. In addition, the lack of CTL immunity should remove any selective pressure and therefore potential advantage, although treatment such as reduced immunosuppression or EBV-specific CTL transfusion could re-establish this evolutionary force. No data on this was available for the PTLD patients in this study.

In two (potentially three) cases of the IC data set, there was clear evidence for co-infection with another EBV strain. Co-infections can either occur during primary infection, i.e. due to transmission of multiple strains, or later acquisition of another virus. One could imagine that new infections might establish themselves easily due to the lack of immune control in immunocompromised patients.

One risk factor of PTLD is EBV seronegativity pre-transplantation (Gulley and Tang, 2010). This is most likely to be the case in children, which means that the primary infection results from the transplant. In the TB data set consisting of longitudinal blood samples, co-infection was not detected.

In some cases, it is hard to differentiate between genetic drift and co-infection. On the one hand, genetic drift would leads to random changes across the genome, while co-infection likely shows additional variants at sites known to vary between isolates. Additionally, reconstructing haplotypes could help disentangling this. However, longitudinal

data is usually necessary for this, as most approaches are cluster-based using frequency changes of variants over time. A combination of long read sequencing of longitudinal samples would therefore be the best strategy to resolve questions about genetic-drift versus co-infection.

Co-infections of EBV, both in terms of co-infections with type 1 and 2, as well as with different strains of a single type have been described before. Previously, EBV infections with multiple strains have been analysed using PCR amplification of polymorphic genes in the setting of HIV-positive, T cell immunocompromised individuals (Yao et al., 1996; Yao et al., 1998) and in immunocompetent individuals (Srivastava et al., 2000; Sitki-Green, Covington, and Raab-Traub, 2003). Moreover, heteroduplex tracking assays (HTA) were being used to describe multiple infection in healthy carriers as well as infectious mononucleosis (Sitki-Green, Covington, and Raab-Traub, 2003; Tierney et al., 2006; Sitki-Green et al., 2004; Kwok et al., 2015). However, there is very little report on minority variants detected via NGS in the literature. One study analysed *EBNA1* variation in MS patients (Tschochner et al., 2016), and another study looked broadly at the intrahost nucleotide diversity in IM patients (Renzette et al., 2014). To our knowledge this is therefore the first study that looked at intrahost variation on the whole genome level.

5.3.2 Primary EBV infection

Whole genomes from longitudinal blood samples of Japanese paediatric patients with IM were sequenced. Due to low viral titres after the onset of immune control, genomes could rarely be recovered from the second time point. For three patients, however, it was possible to obtain longitudinal genomic data. In these patients, the two time points were only a few days apart, yet a few consensus level changes could be observed. Additionally, a number of minority variants was found in most samples where depth was high enough. One patient sample, P12-1026, displayed a higher number of those. But it is not possible to determine, from these data, whether this is due to accumulation of variation of the infecting strain, or whether it might be co-infection with another strain due to missing data (especially depth to reliably call minority variants) and only a single time point. Intrahost nucleotide diversity and the fairly low number of variants, would argue against a different distinct virus. On the other hand, most variants fall within regions that are highly diverse between strains. The sample was taken at a relatively early time point after infection (15 days after onset of disease), suggesting either a surprisingly high mutation rate within this patient or primary transmission of an already more polymorphic pool of virus. As mentioned earlier, an approach to differentiate between drift and co-infection would be long-read data in combination with high depth sequencing of longitudinal data, as this would enable the reconstruction of haplotypes. Moreover, one can consider whether minor variants occur at sites, that show lineages (in other words often differ between viruses). This might be confounded, however, by a sampling effect if coverage and depth are as low as in this case.

One patient displayed three longitudinal consensus level polymorphisms located in the ORF *LF2*. Additionally, five genomes out of eight, for which minority variants could

be detected, had polymorphic sites in the same ORF. Protein LF2 binds Rta (encoded by *BRLF1*), one of the two viral transactivators responsible for the switch to lytic replication. LF2 inhibits Rta pomotor activation and alters it subcellular localisation, thereby blocking EBV replication in cells (Heilmann, Calderwood, and Johannsen, 2010). The ORF was affected in both fast (P1, P6) and slow regressing patients (P4, P11, P12), but none was nonsynonymous. No sites under positive selection were detected in LF2 in the data set of section 4.2.3, but this did not include the new Japanese IM sequences presented here. However, there is only little variation present on the nucleotide consensus level, and barely any on the amino acid consensus level, even with these genomes included. Consequently, a ML tree was very poorly supported. Why there is variation on this locus in particular in IM patients remains therefore elusive. Its regulatory role in virus replication makes it interesting, though, as this might participate in disease progression.

It is possible, that the diversity found within the patients is underestimated by studying the virus population solely in the blood. One study analysed the intrahost nucleotide diversity in IM patients in both oral washes and blood at acute phase and convalescence (Renzette et al., 2014). They found that the diversity during AIM is lower than during convalescence due to a relatively homogeneous founder virus. Interestingly, diversity was generally higher in oral washes than in blood. Moreover, the diversity increases during convalescence in the oral washes, but remains low in B cells.

Another study described virus strain variation within different compartments (saliva, PBMCs and plasma) in children with IM and asymptomatic primary infection based on HTA (Kwok et al., 2015). They found that there is a constant interchange of viruses between circulating B cells and epithelial cells, but that there is a discordance between plasma and saliva/PBMCs. This would suggest another undefined reservoir of EBV (Kwok et al., 2015).

Kwok's implication of a unknown EBV reservoir shedding into the plasma is also very interesting. In the context of this study, the plasma compartment would also be included, as DNA has been isolated from whole blood, i.e. both the PBMC and plasma fraction, and therefore also whatever might be shed into the plasma. This reservoir could be, for example, B cells residing in lymphoid organs or peripheral tissue.

5.3.3 Comparison of diversity of EBV infection in immunocomporomised and immunocompetent patients

The data sets presented here include blood-derived sequences from paediatric patients from both immunocompromised (data sets TB and IC) and immunocompetent (data set IM) patients. In both sets, there is an additional (even if limited) longitudinal component. This allows a comparison of diversity found between the two cohorts.

In both data sets, the observed variability over time was very low (tables 5.2 and 5.8). However, the time difference between sample points varied greatly between the data sets (days in the IM data set, years in the immunocompromised data set, see tables 5.1 and 5.7). The sample size is extremely small to draw any strong conclusions, but one could argue that the variability over time might be greater in the IM data set, as a few variants

on the consensus level could already be observed within days compared to a similar number after years in the immunocompromised patients. But more samples are needed for a reasonable estimation of the substitution rate during acute infection (Fu, 2001).

This observation of higher variability is supported by the intrahost diversity and by the number of minority variants in those two groups (figure 5.12 and 5.13). Immunocompromised patients showed a lower nucleotide diversity and lower number of minority variants (when excluding mixed infections), even though results were not statistically significant.

A possible explanation for these varying degrees of diversity could be differences in the mutation rate. Depending on the responsible process of virus replication, the mutation rate can vary. In latency, the high fidelity host polymerase is being used for viral DNA replication in the normal process of cell division. This also means the genome is only replicated once during S phase (Adams, 1987). In contrast to that, the viral polymerase is being used during lytic cycle and replication is initiated more than once (Hammerschmidt and Sugden, 1988). While the polymerase does have proof reading function (Tsurumi et al., 1993; Tsurumi, Daikoku, and Nishiyama, 1994), its mutation rate is not known and it might be possible that higher lytic replication leads to greater changes in the viral population than clonal expansion of infected B cells.

Traditionally, IM is associated with latent infection (Niedobitek et al., 1997; Kenney, 2007), with its clinical symptoms being a result of the strong immune response against EBV-infected B cells. But the relatively long incubation period of one month suggests that lytic replication precedes the onset of clinical symptoms (Kenney, 2007). Moreover, during acute infection in IM patients, EBV has been found to lytically replicate in B cells in the blood (Prang et al., 1997).

In immunocompromised/PTLD patients, the reason for the high viral load in blood, i.e. whether it is lytic or latent replication, is controversial, and both explanations are not mutually exclusive. Additionally, the number of genomes per cell could be a factor. In paediatric SOT patients, a group with particularly high viral loads have been found to carry two populations of infected B cells: one infected with 1-2 genomes per cell, and another one infected with 20-30 genomes per cell (Rose et al., 2002). Carriers with lower (but still elevated) VL usually only had 1-2 genomes per cell.

Gärtner et al., 2002 argue that the main mode of replication in immunocompromised patients is latent rather than lytic. Similarly, another group finds no evidence of viral DNA in the plasma nor lytic replication gene expression with a restriction of the high viral load to the memory B cell compartment (Qu et al., 2000; Rose et al., 2001). This is supported by a literature study that has found that circulating EBV levels in transplant recipients have an EBV DNA doubling time which corresponds to the doubling time of lymphocytes undergoing cell division (Funk, Gosert, and Hirsch, 2007; Gulley and Tang, 2010).

In contrast to that, another study found that EBV DNA, which normally resides in B lymphocytes, can be detected in plasma in patients with active EBV infection or EBV-related PTLD, both in encapsidated form as well as naked DNA, indicating both virus

production as well as originating partially degraded DNA from dying cells (Gulley and Tang, 2010). There have also been reports of variable *BZLF1* expression in PTLD cases (Vakiani et al., 2008). One study found a higher ratio of EBV lytic replication in PTLD patients compared to immuncompetent hosts and transplant patients without PTLD, but sample sizes for the PTLD group was very low compared to the other groups (Kroll et al., 2011). In fact, a picture seems to emerge where oncogenicity of EBV is primarily related to latent infection, but that lytic replication can play a role in cancer development. In a humanised mouse model, Zta (*BZLF1*)-knockout infected mice develop tumours less frequently than wild-type infected mice, although both viruses establish long-term latency (Ma et al., 2011). In tumours of human adult PTLD patients, latency II and III were detected in 79 % of cases, and lytic replication in 60 % of cases. The subgroup of patients expressing latency III as well as lytic EBV replication displayed a shorter survival and early-onset PTLD (Gonzalez-Farre et al., 2014).

If lytic replication has a higher mutation rate, it would also explain the emergence of consensus level changes after days as well as the higher number of minority variants in the blood of IM patients, In particular also in P12-1026. The patient is part of slow regression group, perhaps due to a particularly high rate of EBV replication, which in turn would lead to a higher number of minority variants found in the blood.

The notion of lytically replicating virus (especially in epithelial cells in the early phase of IM) playing a role and generating more diversity is supported by the studies discussed earlier, which looked at intrahost variation in different compartments in IM patients (Renzette et al., 2014; Kwok et al., 2015). Especially the findings by Renzette et al., 2014 showed a generally higher diversity in oral washes, in particular during acute phase, suggesting a process which introduces more diversity.

Another factor to be considered in this comparison might be the lack of selection pressure from the immune system in immunocompromised patients. This, however, is also dependent on treatments that reinstate a immune response, such as decreasing the immunosuppressive dose and infusion of EBV-specific CTLs (Heslop, 2009) – information that the data sets presented here are lacking.

5.3.4 Comparison Asian NPC and non-NPC isolates

A whole genome comparison of Asian NPC and Non-NPC samples using PCA showed that the first principal component divides roughly between some NPC and Non-NPC samples, excluding a few intermediate samples of both groups (figure 5.15). The SNPs with the largest contribution are distributed across the genome and lie mostly within coding regions (figure 5.16). Of those SNPs, 271 are nonsynonymous and are also distributed equally across the genome.

The majority of the most affected ORFs are very polymorphic genes encoding antigenic proteins. A number of the nonsynonymous SNPs also affect experimentally described epitopes, indicating that these sites might be important adaptations for immune escape. However, IEDB does not provide information about the HLA allele for which these epitopes have been described (in particular whether they might be presented by

HLA alleles common in China or other Asian countries) which restricts a deeper interpretation.

There is a large number of the NPC/Non-NPC discriminating SNPs falling into the BART region of the genome, from which mRNAs as well as miRNAs are transcribed. The BART miRNAs are highly expressed in latently infected epithelial cells, such as NPC (Cai et al., 2006), which usually display latency II. Another miRNA cluster located in the *BHRF1* locus is expressed in particular in cells of expressing latency III. However, none of the SNPs fall within the miRNAs. This is not overly surprising as miRNAs are usually very evolutionary conserved. They play an important role in the viral life cycle by post-transcriptionally regulating gene expression of various targets involved in regulation of apoptosis, immune evasion and the establishment of latency (Kuzembayeva, Hayes, and Sugden, 2014; Vereide et al., 2014; Kang, Skalsky, and Cullen, 2015; Lin et al., 2015). The differential expression in different tissues rather than sequence variation is likely the more important factor in maintaining the balance between virus and host (Cai et al., 2006; Lin et al., 2015).

It has been shown that there exist at least six different splicing forms of BART mRNAs, and they are assumed to have each an ORF, but native proteins have rarely been shown conclusively *in vivo* (Yamamoto and Iwatsuki, 2012). Ten NPC/Non-NPC discriminating SNPs fall into three exons of the BART mRNAs (V, VA', VB). VA' is part of a minor-splicing form RK-BARF0, which is only rarely detected. On the other hand, VB is part of the BARF0 and RPMS1A splicing form and V part of the RPMS1 and A73 splicing form, all of which have been deteted in various EBV-infected cells (BL, pyothorax-associated lymphoma, LCL, T/NK cell lines) (Yamamoto and Iwatsuki, 2012).

The existence of BARF0 and RK-BARF0 proteins have not been unquestionably shown (Thornburg, Kusano, and Raab-Traub, 2004), but proteins produced from constructs showed the localisation in the nucleus (Kienzle et al., 1999). Additionally, RK-BARF0 could lead to increased levels of LMP1 through sequestering Notch (Kusano and Raab-Traub, 2001; Thornburg, Kusano, and Raab-Traub, 2004). In turn, LMP1 as one of EBV's major oncogenes might contribute to NPC pathogenesis through introduction of morphological and phenotypic changes in epithelial cells (Dawson, Port, and Young, 2012).

RPMS1 is another transcript of potential interest. Two SNPs fall into the exons affecting the RPMS1 transcript, and an additional SNPs is within the *RPMS1* reading frame (G155391A, leading to an amino acid change from Asp to Asn). This specific polymorphism has previously been described as being strongly associated with NPC in southern China (Li et al., 2005; Feng et al., 2015; Cui et al., 2016) and has been proposed to be used in a risk prediction model (Cui et al., 2016). *RPMS1* mRNA is transcribed at high levels in epithelial carcinoma cells (Li et al., 2005; Yamamoto and Iwatsuki, 2012), but could not be detected yet on the protein level in cultured NPC cells or NPC tumour biopsies (Al-Mozaini et al., 2009). However, the SNP has been *in vitro* functionally characterised to decrease RPMS1 protein stability in the NPC-associated variant. In addition, a study found the *RPMS1* mRNA expression is upregulated in association with *BZLF1* mRNA, in Akata and P3HR1 cells (two BL cell lines) when stimulated with PMA, suggesting that

induction of lytic infection might play a role (Yamamoto and Iwatsuki, 2012).

While LMP1 is often seen as an important contributor to NPC pathogenesis and is a very polymorphic protein (Dawson, Port, and Young, 2012), only a single SNP (G335D) affecting LMP1 is contained in the list of NPC/Non-NPC discriminating SNPs. This SNP is located within the cytoplasmic domain, but does not affect any of the C-terminal activating regions or the TRAF protein interaction motif. A clear functional association can therefore not be made in this context. However, this finding also suggests that most variation found in Asian LMP1 could rather be due to geography/population structure than disease association.

One key protein is Zta (encoded by *BZLF1*), the major switch from latency to lytic replication and changes here, either in sequence or expression, might contribute to differences in tropism. Among the discriminating, nonsynonymous sites, 17 affect Zta, two of them (195 and 205) in the DNA-interacting bZIP domain (Chen, Reinke, and Keating, 2011).

Increased lytic replication might also be due to the regulation of gene expression. In the data here, the *BZLF1* Zp-V3 promotor variant is most often observed in NPC cases, compared to the Zp-P variant, but differences were not significant. Zp-V3 was first only associated with type 2 sequences. Later it was also frequently found in Asian samples (Jin et al., 2010) and this variant was most often found in NPC cases here. It was first only found in malignant samples (Gutiérrez et al., 2002), but later also in healthy carriers and IM patients (Tong et al., 2003; Martini et al., 2007). Additionally, a study found it might correlate with severe diseases such as CAEBV (Jin et al., 2010). The other promotor variant found in the data here is Zp-P. It was found to be the dominant variant in non-malignant EBV associated diseases in Chinese children (Jin et al., 2010). The comparison done here is rather small in sample size and does unfortunately not elucidate further whether the potential association between Zp and disease might be real.

Another variation that is arguably associated with NPC development are a 30 bp deletion in *LMP1*. The deletion leads to the loss of ten residues in the transformation effector site 2 (TES2) of *LMP1*. Some studies have suggested a higher oncogenicity of this variant (Ai et al., 2012; Chang et al., 2009; Tao et al., 1998), while others did not show a difference (Fielding et al., 2001; Yeh et al., 1997). The variant seems to be distributed world-wide. While a recent meta study showed an association with NPC development restricted to Asian populations, other studies showed it is frequently occurring in healthy carriers in various populations. Here, all Asian genomes except for one Japanese IM isolate exhibited the deletion, strengthening no specific association with NPC.

Apart from the miRNAs, another group of noncoding RNAs expressed by EBV are the EBERs. There is a significantly different occurance of EBER variants between Asian NPC and Non-NPC samples. These results confirm the association of EB-8m with enedemic NPC cases (Shen et al., 2015).

In general, this PCA-based approach allowed us to generate a list of Asian NPC/Non-NPC discriminating SNPs. It is a step further towards differentiating between geographic

variation and disease associated variation, as it enables to filter out Asian-specific variation also found in NPC-nonendemic regions. A comparison of NPC-endemic strains between samples of NPC patients and healthy controls would obviously be important to further remove more local geographic signatures leading to false positives. However, this approach bears the disadvantage that a healthy control might already be infected with a possibly more oncogenic strain, without having developed NPC yet, calling for appropriate control selection (e.g. healthy elderly).

PCA also has some problematic properties intrinsic to the method. Here, we are interested in identifying variables that differentiate between two groups of sequences. PCA, however, aims to describe the overall variability among individuals. In other words, this variability is comprised of both, within-group as well as between-group variation. A method suggested to overcome this is Discriminant Analysis of Discriminant Components (DAPC). It combines PCA and disriminant analysis (DA), with the latter aiming to summarise the genetic differentiation between groups (Jombart, Devillard, and Balloux, 2010). Moreover, another source of false positives are sites identified due to genetic linkage.

To properly assess the significance of the polymorphisms found independent of the method used, functional studies are needed. While we can data mine the variation data to narrow down on potentially interesting study targets, proper functional characterisation is necessary.

In conclusion, I could not find correlation between previously suggested NPC-associated variations. There are many previous studies which focussed on very specific variations in various populations and clinical settings and associations to pathogenesis seem to be controversial. It is not surprising that data set here cannot contribute significantly to this issue. However, our approach has the advantage of allowing a whole genome comparison approach.

Chapter 6

Conclusion & Future work

The advances in EBV WGS demonstrated here, as well as those shown by other groups, now enables the study of a new range of pathologies associated with EBV. Moreover, previous studies can be reevaluated and repeated without gaps in experimental designs to fully uncover the genomic diversity of EBV while minimising the introduction of bias. By using target enrichment strategies it is possible, given sufficient viral load, to sequence directly from blood, saliva and tumour tissue while retaining minority variant alleles at original frequencies. Consequently, this method can now be applied to virus genomics of acute infection such as IM and CAEBV. However, potential malignancy-associated virus variants may be better identified by sequencing both tumour tissue and the geographic "backdrop" (healthy carriers or primary infection). Sequencing from conditions of low VL titre is still problematic and sequencing from spontaneous LCLs is still the only option in some cases (e.g. MS, where EBV load is not always elevated (Pender and Burrows, 2014) or healthy, asymptomatic carriers).

Comparative genomics and evolutionary analysis are a useful tool to explore a pathogen's vulnerabilities. For example, here we identified potential novel epitopes based on the identification of sites under selection. In the future, it would be important to test the presence of these experimentally and to confirm the use of these tools to predict immune selected sites, as this would help in vaccine design, not just for EBV but also other pathogens.

Many questions regarding EBV genomic diversity remain unanswered. This work and others, most notably Chiara et al., 2016, have elucidated more details about the population structure of EBV, albeit with partly divergent results. While the population structure observed here was best explained with two subpopulations, Chiara et al., 2016 report ten subpopulations. Differences in methods and data sets are likely responsible for this and it is therefore important to undertake an extended, comprehensive analysis a) with the same data as used in Chiara et al., 2016, which incorporated further geographic regions, and b) with additional, carefully selected samples of different ethnicities and geographic regions, ideally with good meta-information of the donors.

For example, intra-subpopulation recombinants from the UK and Australia were described in this work. In Chiara et al., 2016, however, these sequences were assigned to a separate subpopulation, which in a tree of non-admixed representatives clustered more

closely to the Asian subpopulation as defined in their study. Second, within the Western/African subpopulation identified in this work, isolates are in a complex relationship with each other. Chiara et al., 2016 observed that Africa appears to carry the largest number of subpopulations. This was not picked up here as the analysis was restricted to type 1 genomes which had a strong bias for sequences from Kenya. Strategic sampling of more regions and ethnicities would be of great interest. Similarly, there is a bias in available sequences from Asia towards NPC samples from Southern China. While we extended the available genome sequences with genomes of non-malignant background from Japan by eleven unique novel genomes, China and Japan have a partly overlapping culture and history, and more samples from other regions of Asia would be of great interest to elucidate a potential finer population structure on the Asian continent.

EBV's population structure is of interest for several reasons:

First, humans and EBV have a long, shared history. The distinct allele frequencies presented by EBV might reflect distinct strains carried by the human ancestors which colonized the world. They could also be the result of genetic drift and coevolution between virus and host. In this data set, it was observed that the signal of subpopulation was strongest in the set of nonsynonymous sites in LD, i.e. a subset of sites on which selection can act. We therefore propose the hypothesis that EBV's allele frequencies are the result of co-evolutionary adaptations to HLA and recent migration of previously non-overlapping human cultures/populations has led to recombinants between distant genotypes. In future work, this needs to be thoroughly tested. For this, additional information on samples and donor are indispensable, due to recent global migrations and multi-cultural societies in many parts of the world.

Second, it is important to incorporate data on population level allele frequencies to exclude geographic biases in association studies. There is a huge interest in identifying disease-associated alleles in EBV, as these would help to assess patient risk for onset or severity of disease. In order to truly assess the effect of variation, functional studies are necessary. However, effects of population structure (i.e. linkage of non-associated variants) need to be accounted for and previous studies have failed to truly disentangle the relative contribution of geographic variation and disease-association, in particular in pathologies that have geographically divergent incidence rates such as NPC. Furthermore, other factors need to be taken into account to choose appropriate controls. Many studies choose asymptomatic carriers of a similar age distribution. However, given EBV's nature of lifelong latent infection, it is impossible to exclude the development of disease later in life for some of the donors. An alternative option would therefore be the choice of asymptomatic, elderly carriers.

The number of East Asian genomes was expanded by eleven unique novel genomes from Japan. A comparison of Asian EBV genomes derived from NPC and non-NPC led to the identification of a number of sites differentiating between the two groups. In the future, one needs to more accurately distinguish between sites relating to pathogenesis of NPC, and those sites that are simply linked or related to a finer geographic signature. The

analysis illustrates the potential WGS can bring to the field of EBV association studies. Again, more isolates to account for geographic divergence and/or choosing appropriate controls will facilitate this.

Another area of interest regarding EBV's genomics is within host diversity. Studying intra-host diversity could illuminate the diversity in different compartments and the relative contribution of lytic and latent replication. It further allows the understanding of the evolutionary dynamics over time in different clinical settings of EBV infection.

A few studies have aimed to study within host diversity using single gene sequencing or HTA with varying results. Kwok et al., 2015 described the persistence and interchange of viral strains based on LMP1 variation using HTA and made the interesting observation of a potential hidden reservoir of EBV interchanging with plasma but not with PBMCs/saliva. Within-host recombination, however, is a likely scenario and could bias studies such as this, which WGS could overcome.

Here, intra-host diversity on the whole genome level of EBV has been described for the first time. Secondary EBV infections were detected in the case of some immunosuppressed children. In paired tumour- and blood samples of PTLD patients, tumours were found to be completely monoclonal, while blood samples contained a few minor variants. A comparison of paediatric immunosuppressed and immunocompetent patients revealed a slightly higher diversity in immunocompetent children with primary infection, but differences were not significant. It is unclear, however, whether this is due to secondary infections or within host evolution, as this approach was limited by the depth achieved by sequencing EBV from blood. Improvements in EBV sequencing will overcome this.

Lastly, EBV genome sequencing, in this work as well as in other published studies, has been hampered by short read length of the most common NGS platforms such as Illumina, which affected in particular the numerous repeat regions of the genome. However, repeat regions can be functionally important and to date, we are missing all the diversity found there. The application of novel long read sequencers such as the Nanopore MinION will help tackling this issue.

Appendix A

Appendix for Chapter 3

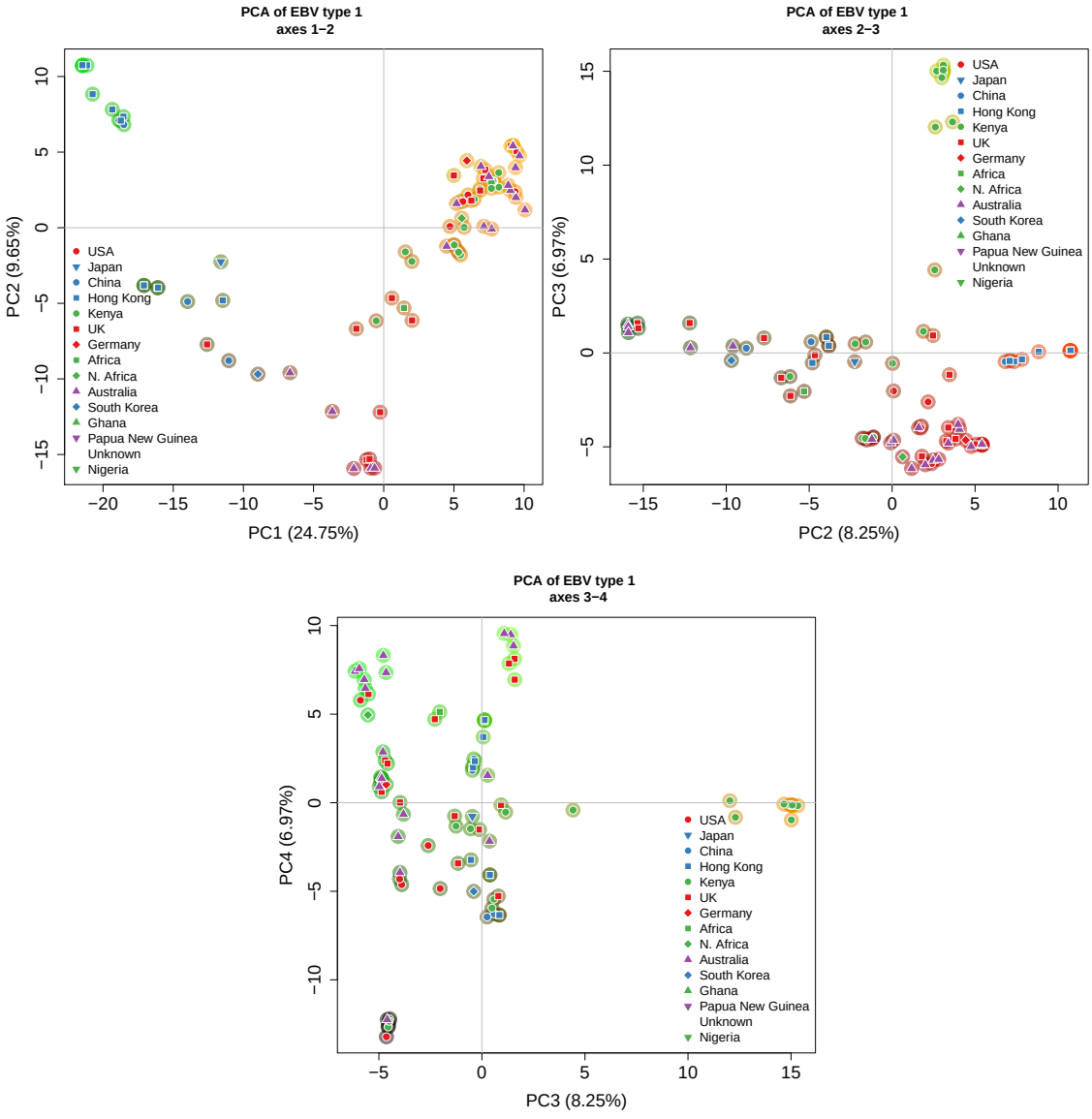


FIGURE A.1: PCA of whole genome sequences for type 1. This is the same plot as in 3.11, but samples are labelled based on their geographic origin.

Appendix B

Appendix for Chapter 4

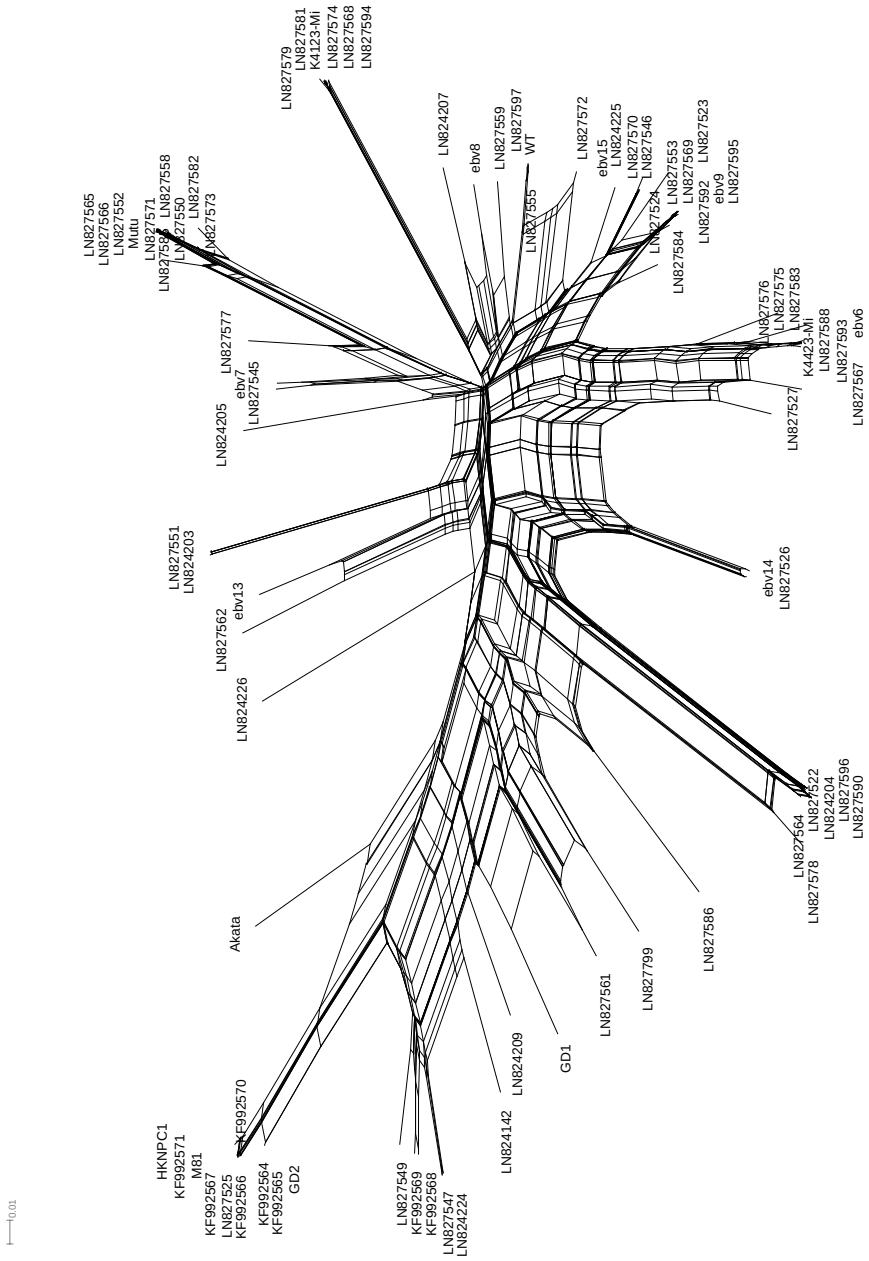


FIGURE B.1: Split network of all sites in LD.

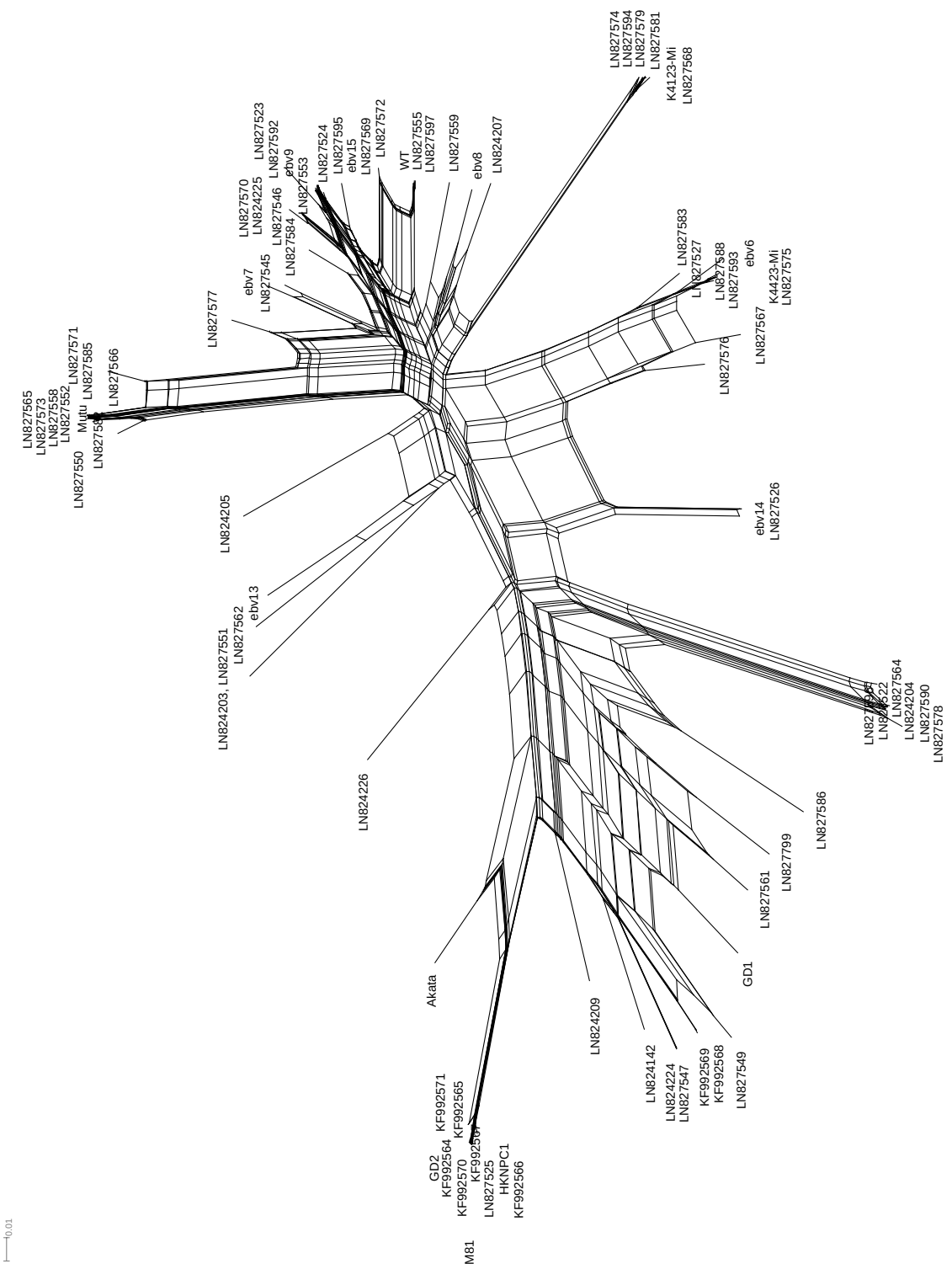


FIGURE B.2: Split network of all nonsynonymous pairs of sites in LD.

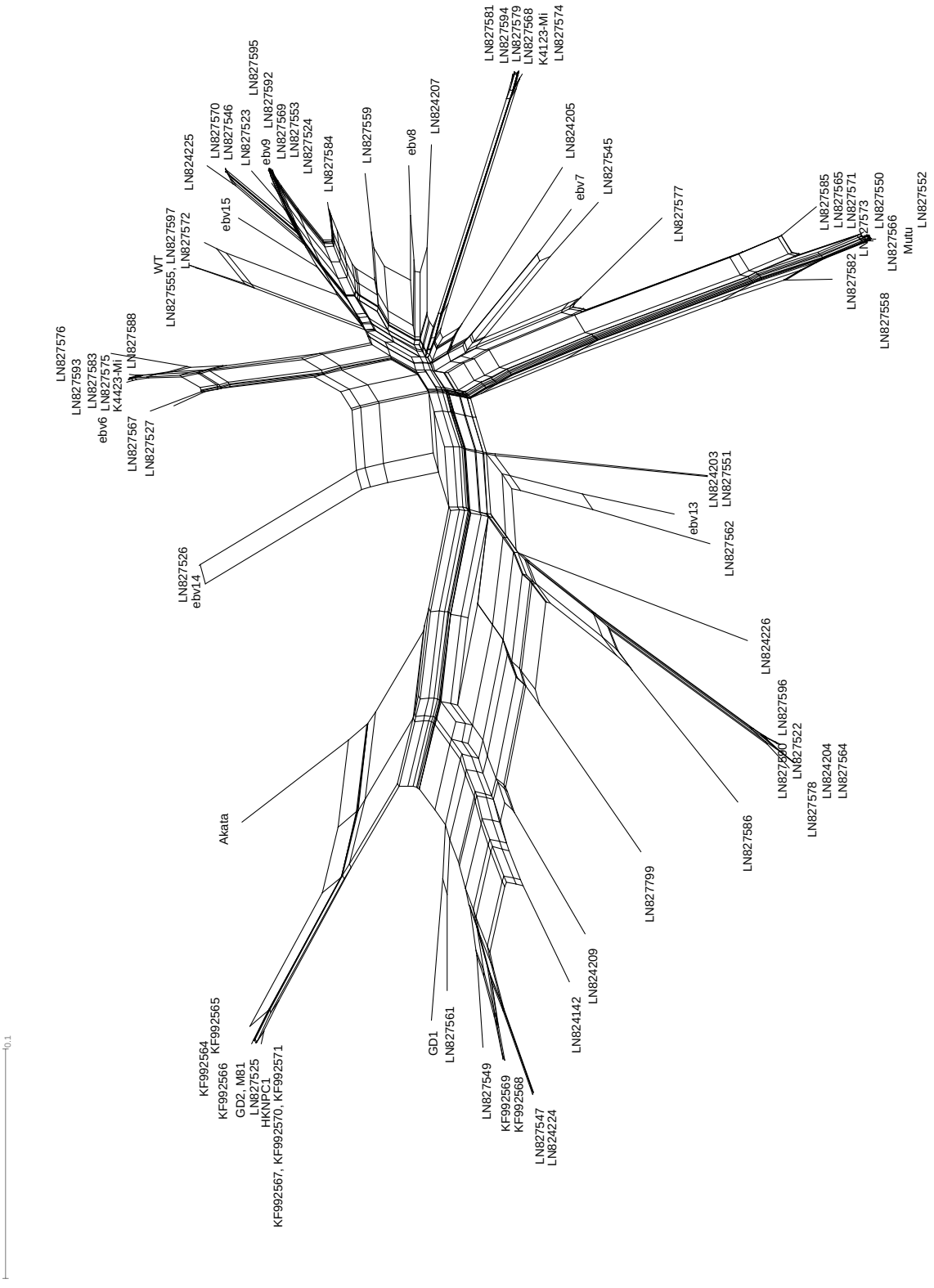


FIGURE B.3: Split network of all synonymous pairs of sites in LD.

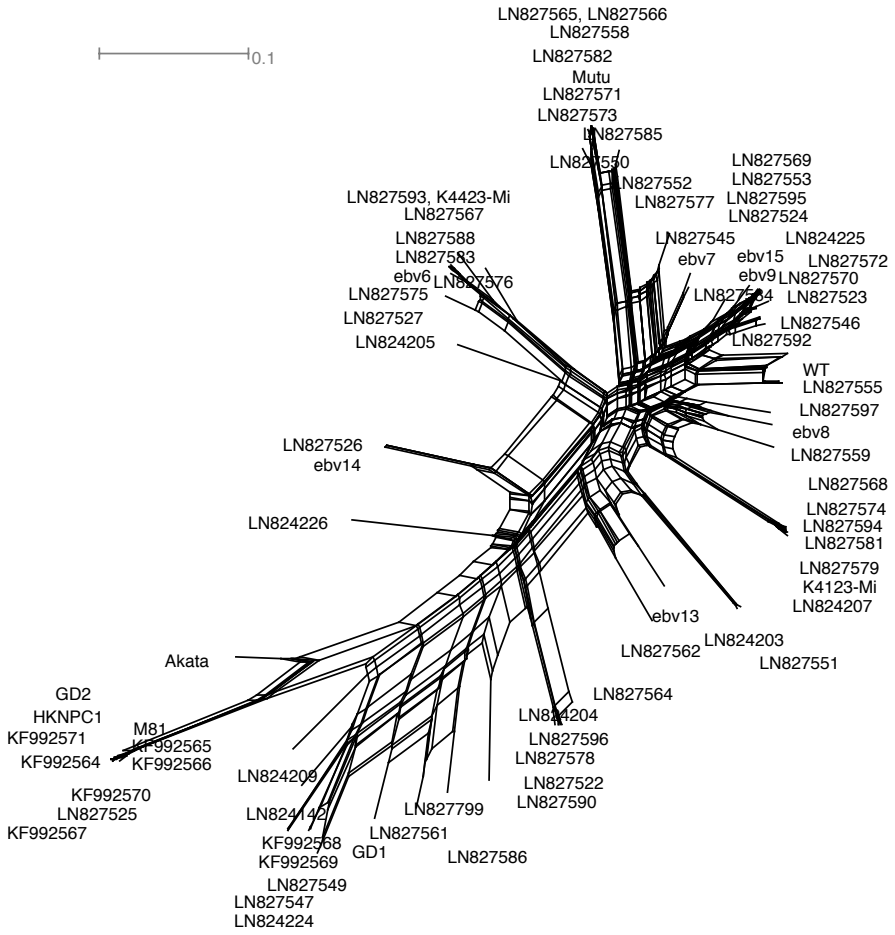


FIGURE B.4: Split network of biallelic sites in LD with a threshold of $p < 0.001$.

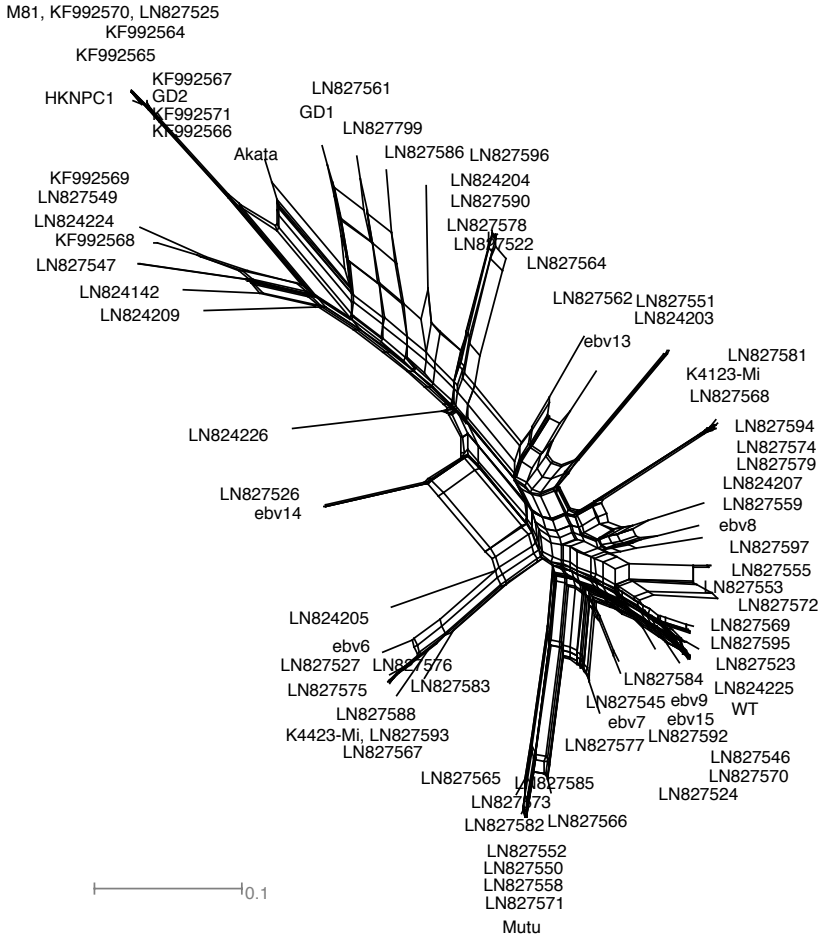


FIGURE B.5: Split network of biallelic sites in LD with a threshold of $p < 0.002$.

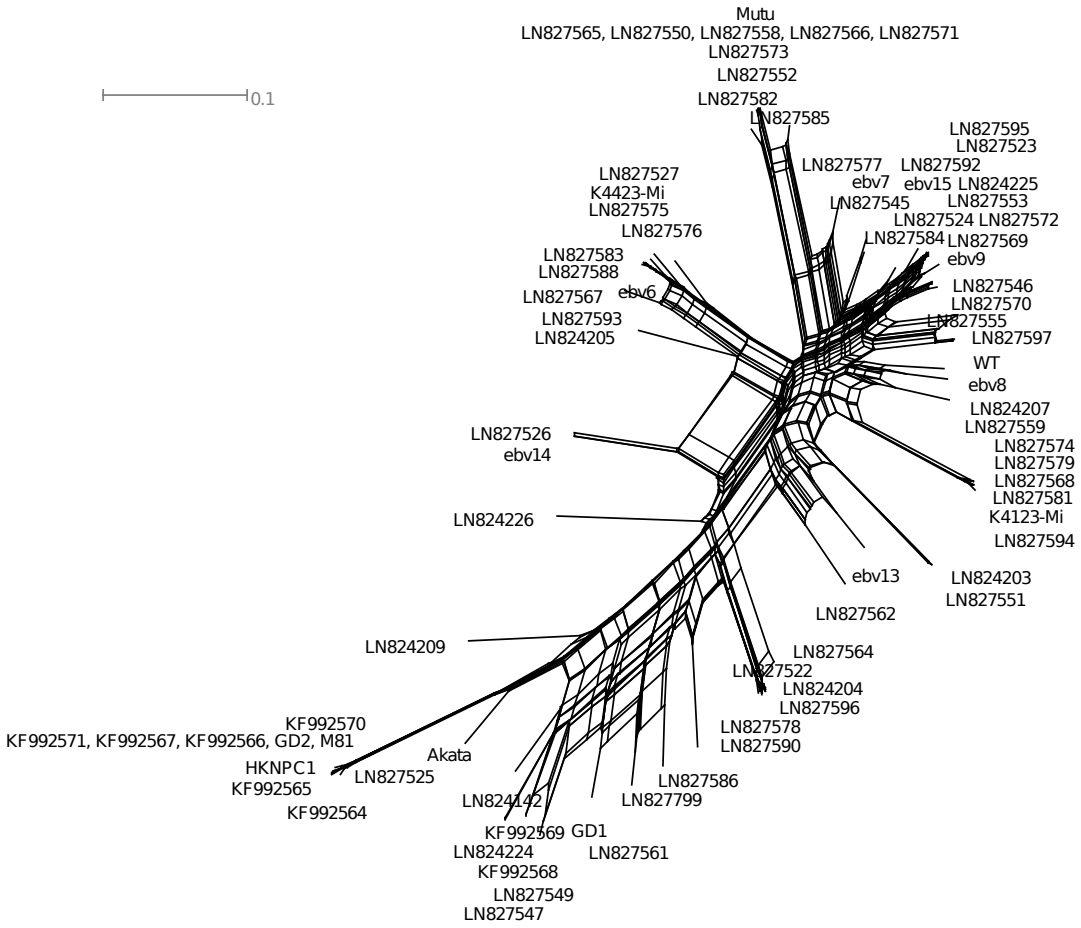
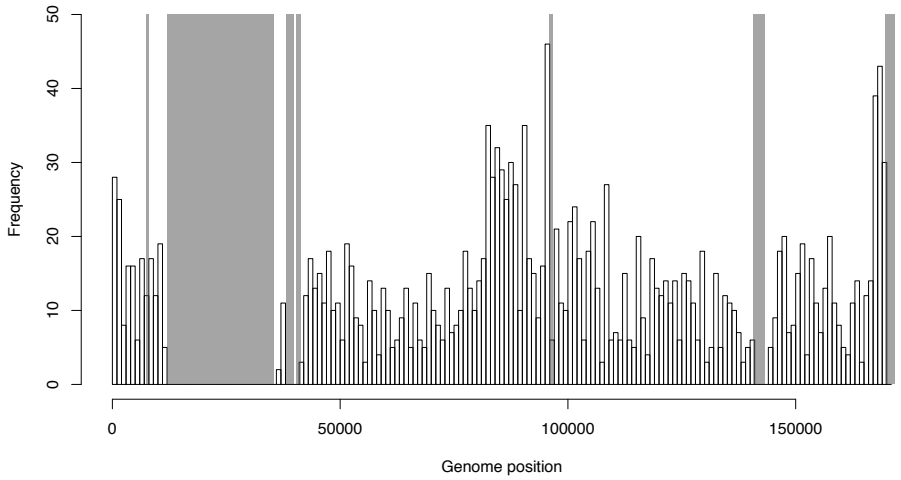
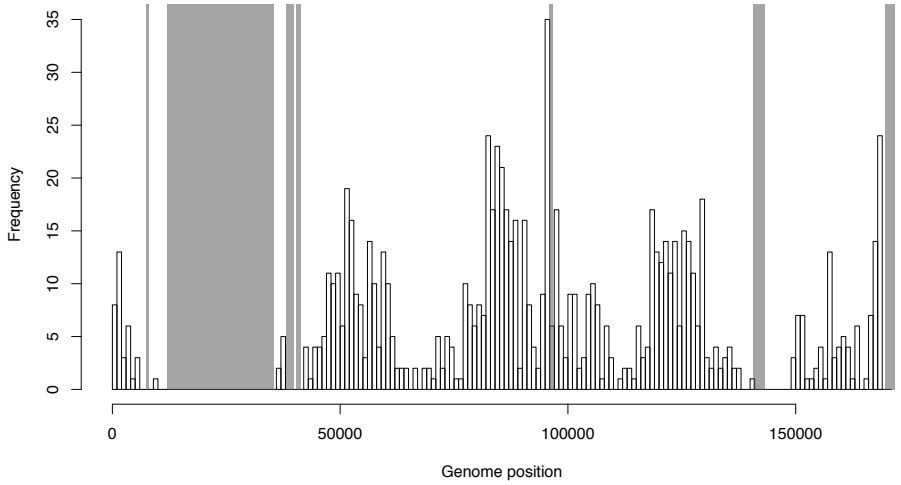


FIGURE B.6: Split network of biallelic sites in LD with a threshold of $p < 5E-05$.



(A) All



(B) Nonsynonymous

FIGURE B.7: Histograms of sites in LD across genome in 1 kb bins. Grey areas mark the repeat regions.

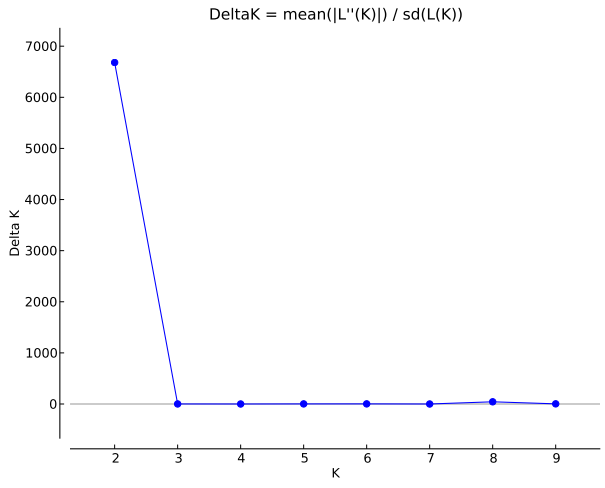


FIGURE B.8: Δk determines the best fitting number of clusters or subpopulations (k) as the k corresponding to the highest value of ΔK (Evanno, Regnaut, and Goudet, 2005).

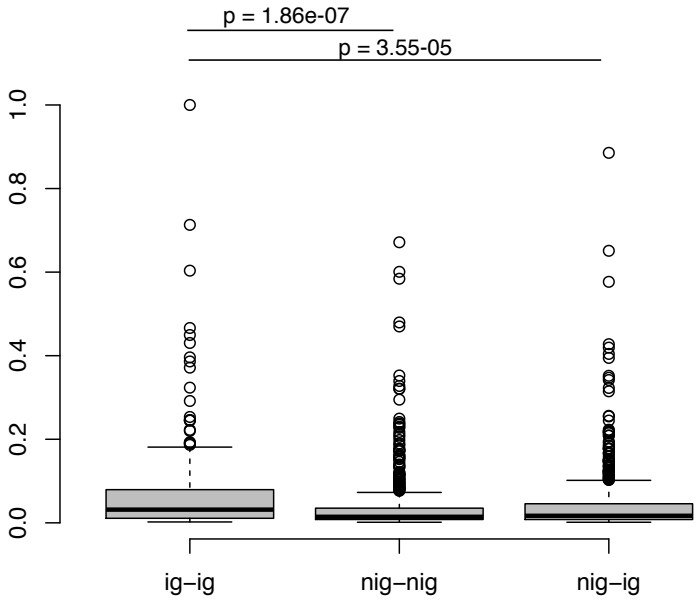


FIGURE B.9: Linkage score (edge weight) between nodes belonging to different gene categories. Differences were tested with Mann-Whitney U test.

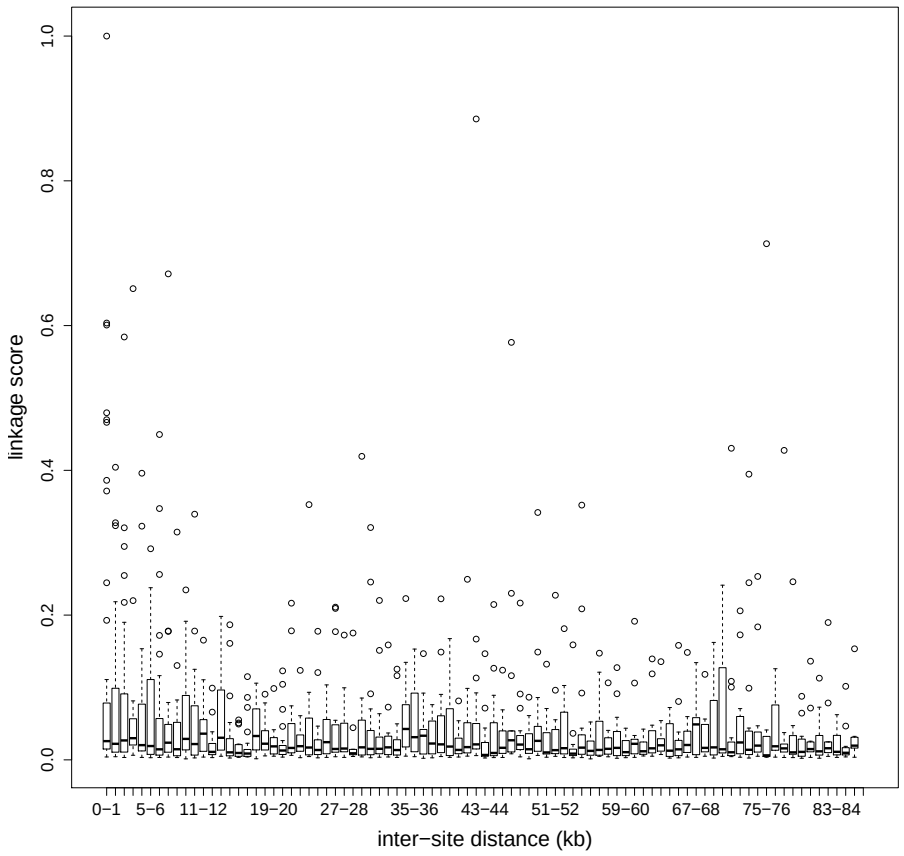


FIGURE B.10: Distribution of linkage scores in the gene network over genomic distance. The distance between two linked genes has been binned into distance classes of 1 kb size.

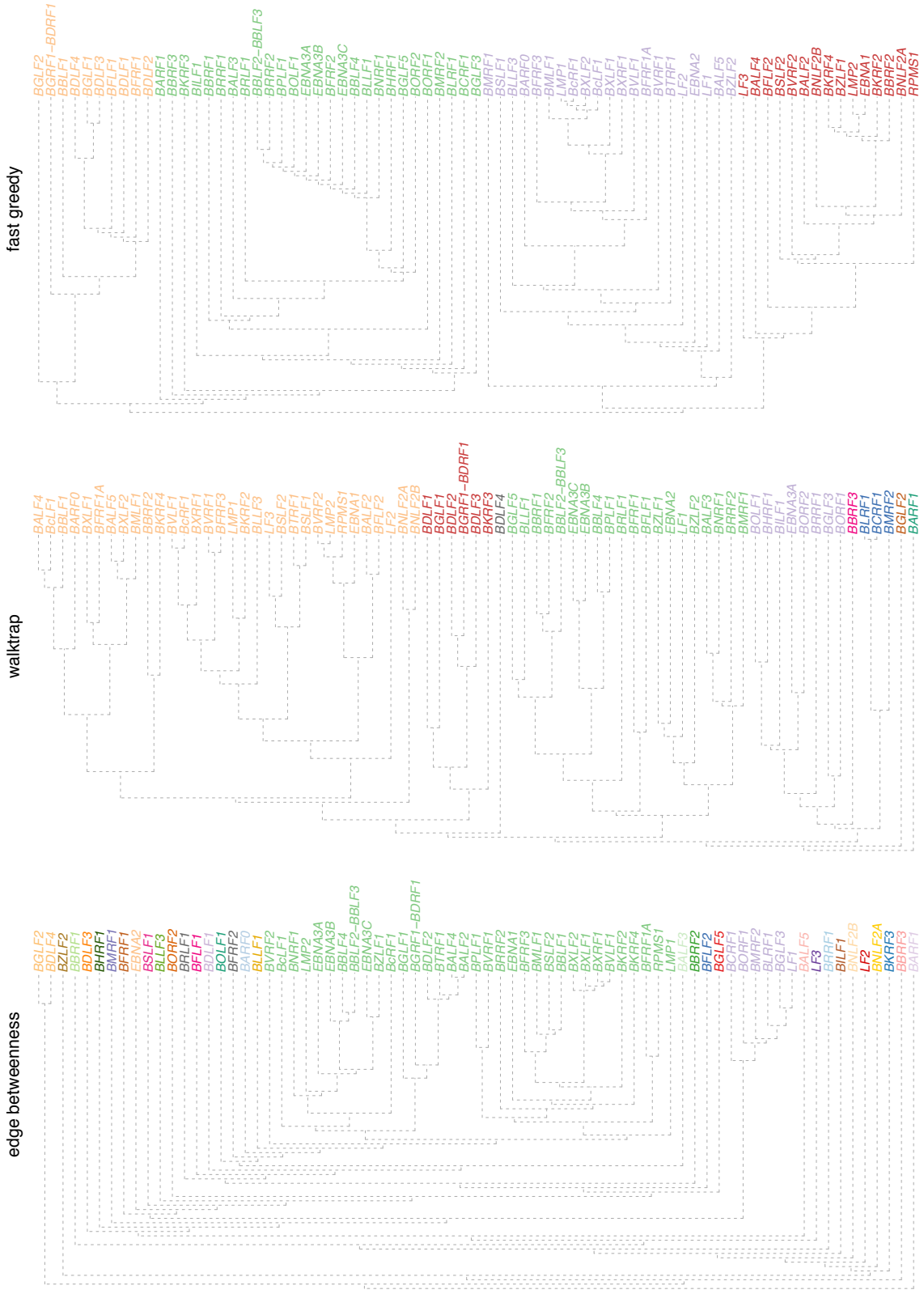
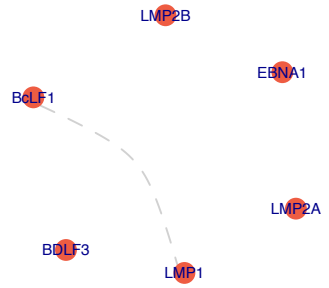
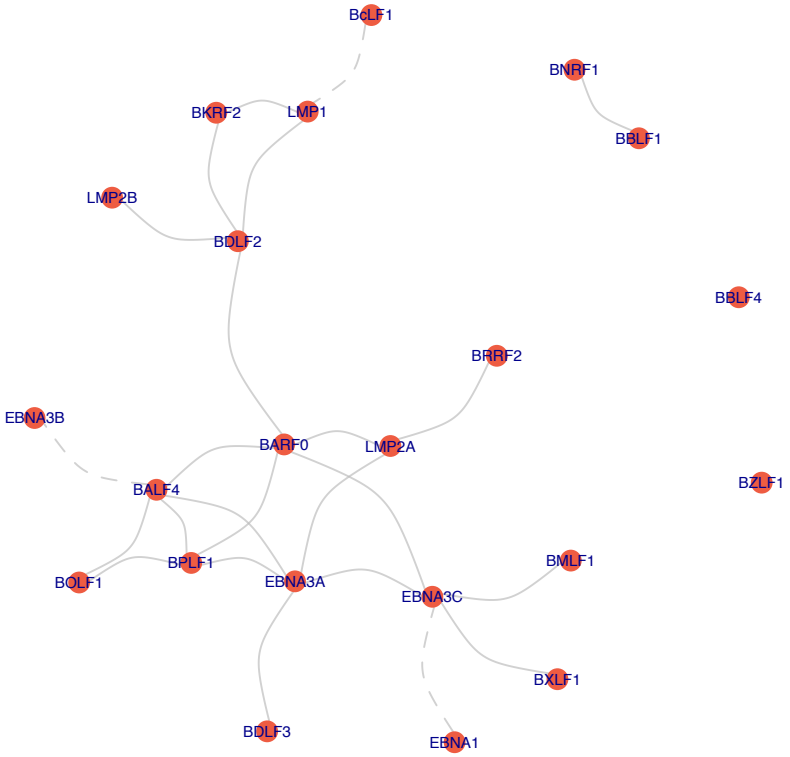


FIGURE B.11: Results of different graph clustering algorithms.



(A) PPI between closest linked nodes according to hierarchical clustering of the gene network (grey shaded box in figure 4.10).



(B) PPI between top 25 Eigenvector centrality ranked nodes of the gene network.

FIGURE B.12: Protein-protein interactions (PPI) between top ranked nodes of the gene network according to different approaches where data is available. Interaction data is from Calderwood et al., 2007 and Fossum et al., 2009 based on Yeast two hybrid screenings. Low confidence interactions (Calderwood et al., 2007) are shown as dashed lines.

	pos aa	pos nt	pos genome	residue	substi-tution	M1a-M2a	M7-M8
BALF2	1065	3195	161117	E	G	**	**
	1093	3279	161033	S	G		*
BALF3	5	15	160893	V	I	**	**
	62	186	160722	Q	R	**	**
	127	381	160527	T	A	**	**
BBRF3	15	45	106894	V	F		*
BLLF1	71	213	79652	H	Y/R	**	**
	72	216	79649	T	M	*	**
	201	603	79262	E	Q	**	**
	755	2265	77600	Q	K		*
BNLF2B	6	18	166818	P	K/S	**	**
	72	216	166620	Q	K	**	**
BNRF1	459	1377	3113	G	R	*	**
	500	1500	3236	S	G/N	*	*
	552	1656	3392	G	S	*	*
	739	2217	3953	C	F	*	*
BOLF1	492	1476	61475	A	V/T		*
	1179	3537	59414	R	H		*
BPLF1	12	36	59203	T	P	**	**
	625	1881	57358	P	S		*
	636	1914	57325	P	S/L	*	*
	648	1950	57289	D	E/Y	**	**
	654	1968	57271	S	A	*	*
	76	2394	56845	S	A	**	**
	1528	4584	54655	A	V	*	*
	1535	4605	54634	D	E	**	
	1808	5424	53815	A	S	*	
	1879	5637	53602	V	F	*	
	2087	6261	52978	R	K	*	
	2246	6738	52501	N	D	**	
	2616	7848	51391	K	E	**	
	2657	7971	51268	V	A	**	
	2696	8088	51151	S	T/R	**	
	2736	8208	51031	H	N	**	
	2765	8295	50944	P	S/L	*	
	2869	8607	50632	F	L	*	
	2895	8685	50554	R	S	**	
	2897	8691	50548	Q	K	*	
	3000	9000	50239	R	Q	**	
BRLF1	290	870	92025	A	S/D		**
	377	1131	91764	A	E	**	**
	479	1437	91458	V	I		*
	489	1467	91428	Q	R/K	**	**
	542	1626	91269	S	N		*
BRRF2	184	552	94566	L	P	**	**

	pos aa	pos nt	pos genome	residue	substi-tution	M1a-M2a	M7-M8
	260	780	94794	R	Q	*	*
	285	855	94869	Q	K	**	**
	313	939	94953	H	R	**	**
	323	969	94983	R	Q	**	**
	325	975	94989	S	L	**	**
BSLF1	293	879	73714	M	V 8	*	*
	404	1212	73381	R	Q	**	**
BVRF2	482	1446	137084	S	P	*	**
	540	1620	137258	A	T/V		*
BZLF1	146	438	90429	V	A		*
	205	615	90042	A	S	**	**
	206	618	90039	A	S/T	**	**
BcRF1	738	2214	127391	K	R	*	**
	745	2235	127412	C	R/H	**	**
BdRF1	222	666	137084	S	P	*	**
EBNA1	411	1233	96895	E	D/Q/G 8	**	**
	429	1287	96949	V	M	*	*
	487	1461	97123	A	T/L/V	**	**
	499	1497	97159	D	E	**	**
	500	1500	97162	E	D		*
	502	1506	97168	T	N	*	*
	524	1572	97234	T	I	*	*
	582	1746	97408	L	F	**	**
	584	1752	97414	M	L	**	**
	585	1755	97417	T	P/I	**	**
	588	1764	97426	A	P		*
	594	1782	97444	R	K	**	**
	595	1785	97447	V	A	*	*
EBNA2	23	69	36285	L	V/R	**	**
	42	126	36342	D	G/E		*
	163	489	36705	R	M/V		*
	187	561	36777	R	S/K		*
	194	582	36798	L	F/P	**	**
	195	585	36801	M	T/I		*
	315	945	37161	L	S		*
	475	1425	37641	Y	F/H	**	**
	477	1431	37647	E	G/V		*
EBNA3A	189	567	80611	T	M/L	**	**
	190	570	80614	T	A/N	*	*
	219	657	80701	L	P	*	*
	293	879	80923	S	N/G	**	**
	333	999	81043	I	L/Q	*	*
	459	1377	81421	P	T/S/R	*	*
	561	1683	81727	I	F	**	**
	620	1860	81904	P	T	**	**

	pos aa	pos nt	pos genome	residue	substi-tution	M1a-M2a	M7-M8
	647	1941	81985	F	S/E	**	**
	648	1944	81988	S	Q	*	*
	655	1965	82009	R	H/Q	**	**
	656	1968	82012	A	T	*	*
	675	2025	82069	I	T		*
	681	2043	82087	V	A/M	**	**
	733	2199	82243	Q	R	**	**
	764	2292	82336	N	T	**	**
	807	2421	82465	A	V/D	**	**
	814	2442	82486	G	A/V/T/D	**	**
EBNA3B	33	99	83164	T	K/Q	**	**
	173	519	83663	C	H/Y	*	*
	212	636	83780	T	M/L/P/R	**	**
	462	1386	84530	Q	H	**	**
	489	1467	84611	A	T	**	**
	556	1668	84812	S	L	*	*
	611	1833	84977	T	M	*	*
	615	1845	84989	R	Q/W	*	*
	848	2544	85688	A	E	*	*
	900	2700	85844	G	S	**	**
	901	2703	85847	Q	R/K	**	**
EBNA3C	21	63	86146	N	D		*
	44	132	86215	R	G	*	*
	51	153	86236	Y	D	*	*
	104	312	86395	T	A/P	*	*
	107	321	86404	T	V/I	**	**
	141	423	86581	I	V	*	*
	162	486	86644	A	V	*	*
	215	645	86803	A	G/E	*	*
	277	831	86989	L	M	*	*
	357	1071	87229	G	V/I	**	**
	821	2463	88621	T	A	*	*
	978	2934	89092	A	V/S	**	**
LF1	8	24	151670	Q	H	**	**
	234	702	150992	G	E	**	**
LMP1	2	6	169010	E	D 8	*	*
	3	9	169007	H	L/R	**	**
	18	54	168962	G	R/Q	**	**
	25	75	168941	L	I/R/S	**	**
	26	78	168938	G	R/L/V/A	**	**
	43	129	168887	V	I/T/L	**	**
	46	138	168878	D	N	**	**
	63	189	168827	I	L/V/M	**	**
	84	252	168764	C	G/A		*
	101	303	168634	H	Q/R/N/S	**	**

	pos aa	pos nt	pos genome	residue	substi-tution	M1a-M2a	M7-M8
	115	345	168490	G	A	**	**
	210	630	168205	D	A/N/Y		*
	212	636	168199	G	S/T/R/A/C	**	**
	214	642	168193	E	Q/D/H	**	**
	229	687	168148	S	T	**	**
	318	954	167881	G	K	*	**
	322	966	167869	Q	N/E/T/D/K	**	**
	331	993	167842	G	Q/R		*
RPMS1	3	9	150332	G	E		*
	18	54	155296	Y	C		*
	20	60	155302	A	V		*
	24	72	155314	P	H		*
	29	87	155329	G	E		*
	49	147	155389	P	L		*
	50	150	155392	P	L		*
	51	153	155395	D	N		*
	55	165	155407	R	Q		*
	91	273	155515	G	C		*
	98	294	155536	S	N		*
	99	297	155539	C	Y		*
	103	309	155551	R	K		*

TABLE B.1: Positively selected sites. Protein, gene and genomic positions refer to the WT genome. *: 95 %; **: 99 % BEB posterior probabilities for codeml (Yang, 2007) analysis.

	π	D	p
A73	0.002329303	-1.360452592	0.162943913
BALF1	0.002172382	-2.156903008	0.010255326
BALF2	0.003958144	-1.166916023	0.245112234
BALF3	0.005440788	-0.282738891	0.822481383
		-1.283155305	0.193592065
		-0.554614361	0.619247514
BALF4	0.003982123	-1.620375316	0.083352211
		-0.757734698	0.478639791
BALF5	0.003862471	-1.387852965	0.152929779
		-1.091755985	0.281884668
BaRF1	0.004159379	-0.852229463	0.418068776
BBLF1	0.005395712	0.529843774	0.585831948
BBLF2-BBLF3	0.007003594	-1.54457441	0.103142588
		0.12412682	0.867043136
BBLF4	0.006308038	-1.134572916	0.260438429
		-1.552405138	0.100908706
BBRF1	0.005088512	-1.034516774	0.31197938
BBRF2	0.003982382	-1.187865443	0.235047079
BBRF3	0.003989965	0.011574032	0.951476534
		-1.057628583	0.299634165
BcLF1	0.004800532	-0.707323303	0.512328142
		-0.148401478	0.925960444
BcRF1	0.006336994	-0.908039043	0.384078966
BCRF1	0.001504379	-1.569045988	0.096537261
BDLF1	0.004525912	-0.585375845	0.597228165
BDLF2	0.003547349	-1.44050082	0.134853182
		-1.210385825	0.224822587
BDLF3	0.01107506	-0.699332021	0.517747888
BDLF4	0.006214425	-0.515552792	0.647738847
BdRF1	0.004537489	-1.782476748	0.050035382
BFLF1	0.005050547	-0.160185248	0.916977791
		-1.242733317	0.210437066
BFLF2	0.005799323	-0.840069657	0.425664063
		-0.417356135	0.720552785
BFRF1	0.005059306	-1.406697385	0.146269462
BFRF1A	0.006959805	-0.617655365	0.574333688
BFRF2	0.005654338	-0.905986569	0.385307256
		-1.24488196	0.209641178
BFRF3	0.004651978	-1.585799533	0.092270581
BGLF1	0.006025708	-0.827731189	0.433432771
BGLF2	0.002436647	-1.07447274	0.290801598
BGLF3.5	0.000483985	-1.6546416	0.075331389
BGLF3	0.004181198	-0.950732022	0.358870787
BGLF4	0.003144808	-0.695256273	0.520519161
BGLF5	0.003934926	-0.334107171	0.783520763
		-0.684071939	0.52815753
BGRF1-BDRF1	0.002682059	-1.612866189	0.085247037
BHRF1	0.004126864	-1.662678669	0.073528615
BILF1	0.003533138	-0.960223079	0.353394413
BILF2	0.001504564	-1.885021802	0.034580557

	π	D	p
BKRF2	0.004988368	-0.681321908	0.530041978
BKRF3	0.003856782	-0.308373351	0.803150941
BKRF4	0.004251064	-0.616579383	0.575092715
BLLF1	0.008725829	-1.449624963	0.131783289
		-1.783247008	0.049903997
BLLF2	0.008901829	-1.124798555	0.265252778
BLLF3	0.002298161	-1.758036707	0.054330444
BLRF1	0.001450522	-1.874863541	0.0359353
BLRF2	0.005975062	-1.3362279	0.172124051
BMLF1	0.004512425	-1.470424818	0.12510468
BMRF1	0.002441708	-1.713699721	0.062762682
BMRF2	0.004995081	-0.804638253	0.44813712
BNLF2A	0.006826267	-0.750155686	0.483646036
BNLF2B	0.006191014	-0.702023451	0.515917736
BNRF1	0.004680507	-1.694207087	0.066738981
		-1.43763078	0.135735207
BOLF1	0.006275237	-1.554532805	0.100808086
		-1.745616232	0.057001133
BORF1	0.005132135	-1.592501182	0.090414133
BORF2	0.004475411	-0.880043076	0.400927619
		-1.22445008	0.218463817
BPLF1	0.005954189	-2.921785888	$6.09\mathrm{e}{-7}$
		-1.177893155	0.240083623
		-0.973586625	0.346029218
BRLF1	0.00668177	0.185353287	0.82184647
		-0.485429273	0.669873692
BRRF1	0.002789568	-1.748626298	0.056050325
BRRF2	0.013771688	-0.430661888	0.710581847
BSLF1	0.004557408	-0.914783957	0.380018659
		-1.382876701	0.15479013
BSLF2-BMLF1	0.004196182	-1.504846625	0.114535161
BSRF1	0.002817085	-1.598448962	0.088773523
BTRF1	0.004190326	-0.811231796	0.443937791
BVLF1	0.005363463	-0.892004329	0.39365711
BVRF1	0.004989001	-1.918619795	0.030359291
		-1.055924908	0.300535287
BVRF2	0.003985609	-1.719471316	0.061617257
BXLF1	0.003299354	-0.747446779	0.485440465
		-1.038743364	0.309702329
BXLF2	0.004038546	-0.820018042	0.43834254
		-0.914154299	0.380393498
BXRF1	0.005800498	0.315203327	0.728974821
BZLF1	0.013921389	0.299317891	0.740110857
BZLF2	0.005882372	-0.470317951	0.681048832
EBNA1	0.014437942	-0.525660585	0.640356452
		0.546393416	0.57541632
EBNA2	0.008714703	-1.152551535	0.252817253
EBNA3A	0.009894401	-1.451741547	0.132070918
		-0.985932461	0.339424787
EBNA3B	0.011302704	-1.180669454	0.239295401

	π	D	p
		-1.42880094	0.13967507
EBNA3C	0.009958089	-0.757928833	0.478704815
LF1	0.007702755	-1.21345248	0.22344184
		-1.091858747	0.281889084
LF2	0.003140119	-1.465896288	0.126539932
LMP1	0.030201344	-0.636759965	0.560913111
		-2.574544895	0.000466718
LMP2	0.010775619	-1.322469248	0.177474791
		-1.104772201	0.275267523
RPMS1	0.004129052	-1.567006855	0.096948967

TABLE B.2: Nucleotide diversity for whole ORFs and Tajima's D values with respective p value for ORFs or ORF fragments if recombination breakpoints have been detected as depicted in figure 4.12.

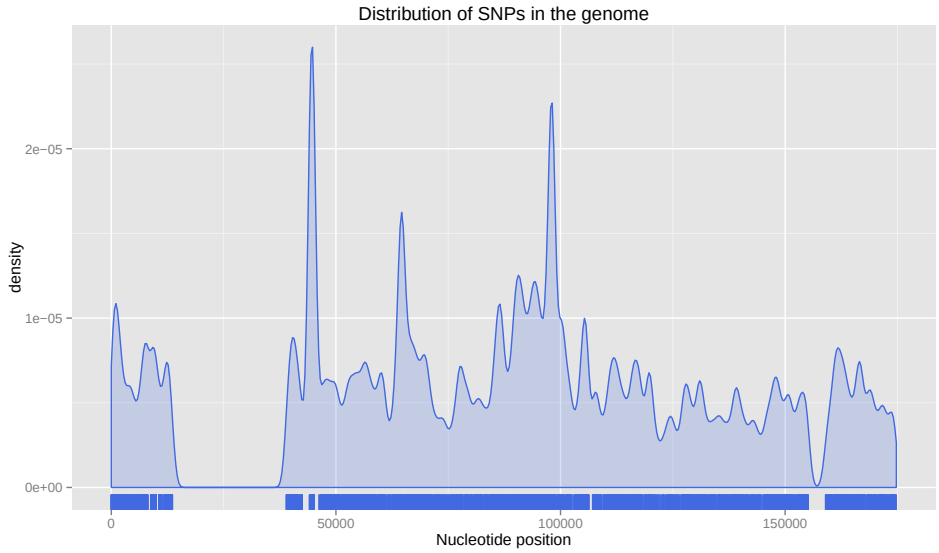
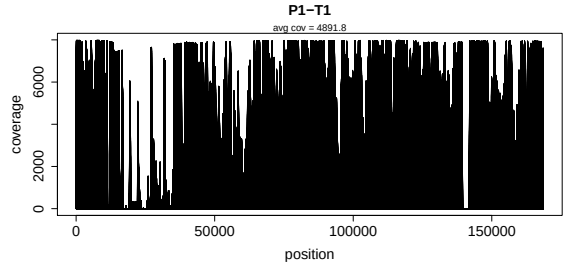


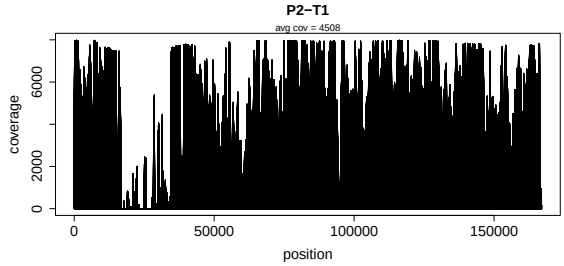
FIGURE B.13: Density of biallelic SNPs across the genome. Note that extremely low values are due to missing data in the repeat regions.

Appendix C

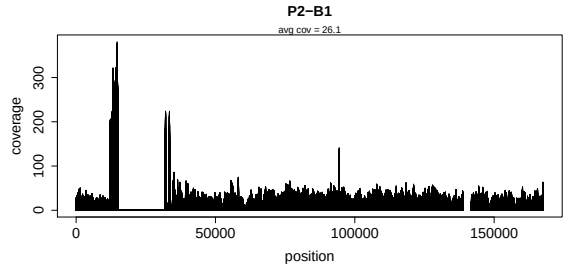
Appendix for Chapter 5



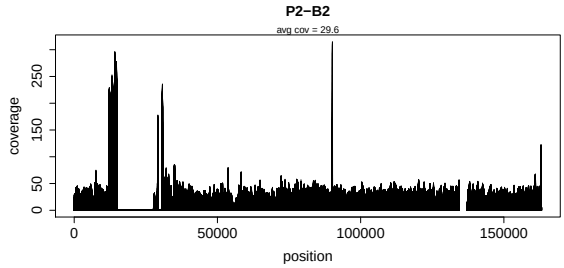
(A) Patient 1, tumour



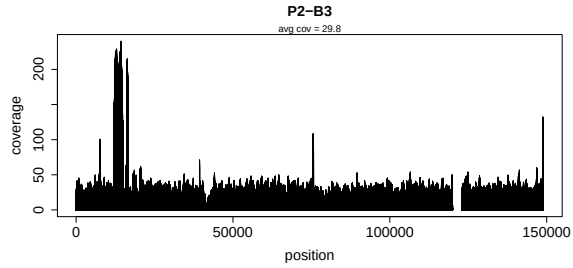
(B) Patient 2, tumour



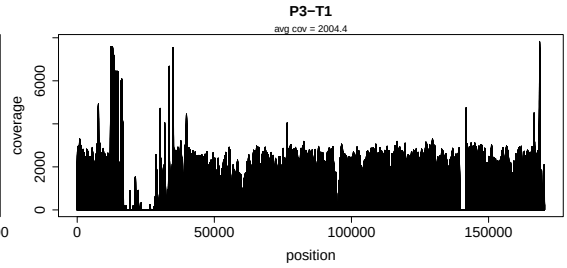
(C) Patient 2, blood 1



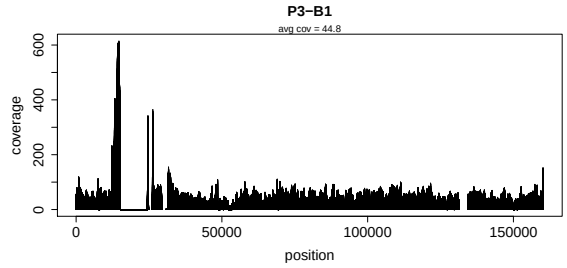
(D) Patient 2, blood 2



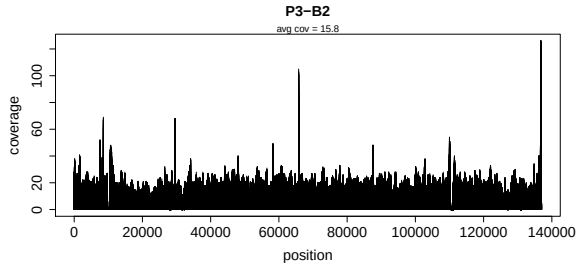
(E) Patient 2, blood 3



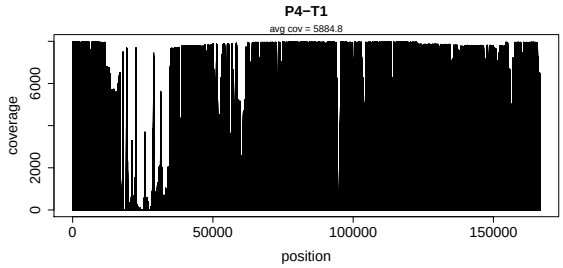
(F) Patient 3, tumour



(G) Patient 3, blood 1

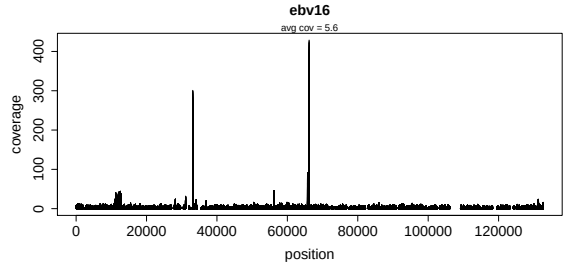


(H) Patient 3, blood 2

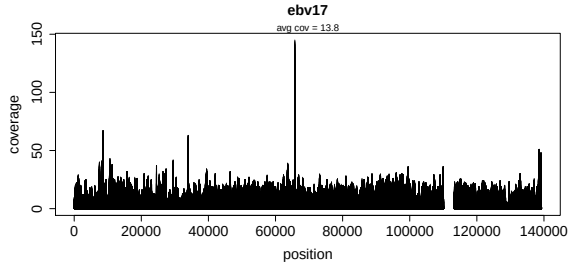


(I) Patient 4, tumour

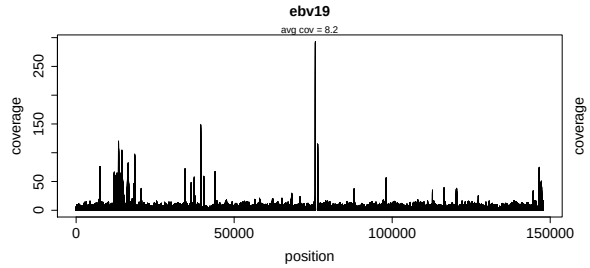
FIGURE C.1: Coverage plots of paired tumour and blood samples after duplicate removal. Mapping is done against the sample consensus sequence directly after assembly (i.e. repeat regions are not masked or considered specifically).



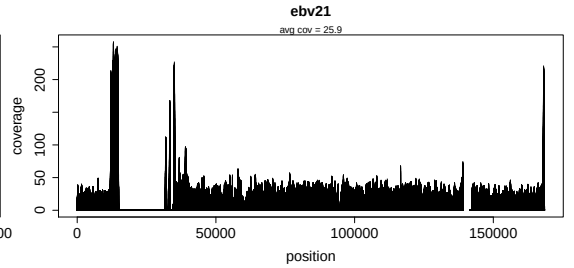
(A) ebv16



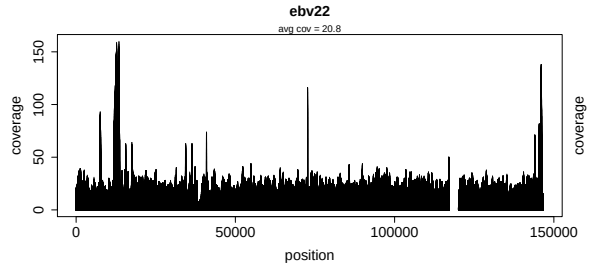
(B) ebv17



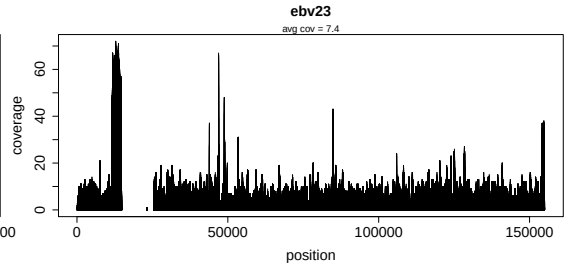
(C) ebv19



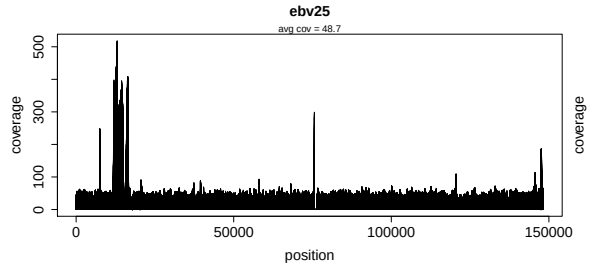
(D) ebv21



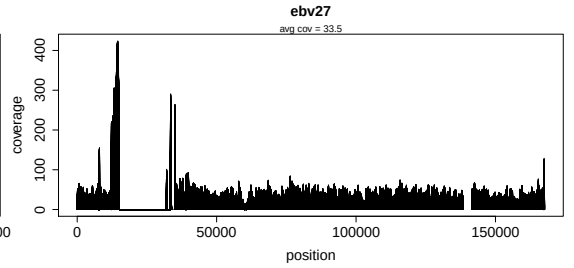
(E) ebv22



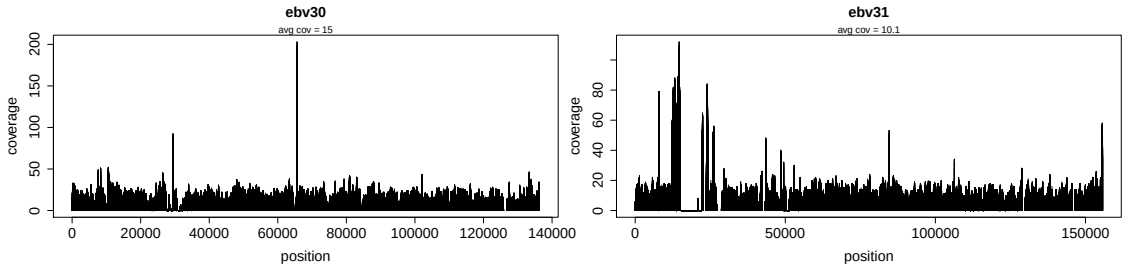
(F) ebv23



(G) ebv25



(H) ebv27

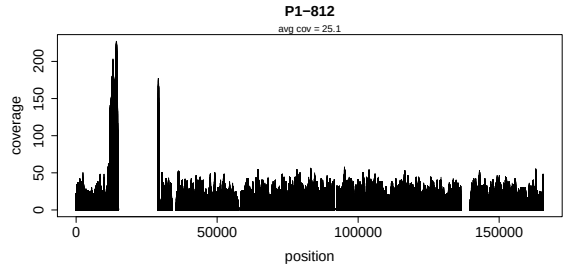


(I) ebv30 (J) ebv31

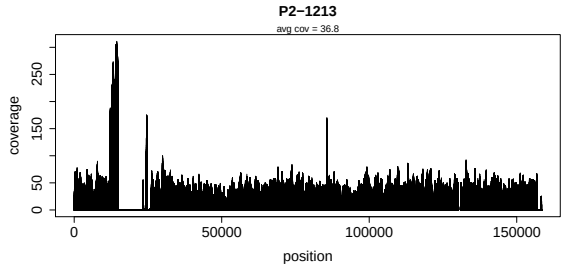
FIGURE C.2: Coverage plots of immunocompromised patient samples after duplicate removal. Mapping is done against the sample consensus sequence directly after assembly (i.e. repeat regions are not masked or considered specifically).

Patient	Sample	Protocol			Sequencing	
		3 µg manual	200 ng manual	200 ng automated	MiSeq	NextSeq
1	P1-812			×	×	×
	P1-833			×	×	
2	P2-1213	×	×	×	×	×
	P2-1246	×	×	×	×	×
3	P3-2670			×	×	×
	P3-2740			×	×	×
4	P4-2274			×	×	×
	P4-2392			×	×	×
5	P5-1294			×	×	×
	P5-1323			×	×	×
6	P6-1751			×	×	×
	P6-1789			×	×	
7	P7-2315			×	×	×
	P7-2634			×	×	×
8	P8-414			×	×	×
	P8-516			×	×	×
9	P9-2631			×	×	×
	P9-2645			×	×	×
10	P10-2187			×	×	×
	P10-2777			×	×	
11	P11-871			×	×	×
	P11-920			×	×	
12	P12-1026			×	×	×
	P12-1078			×	×	

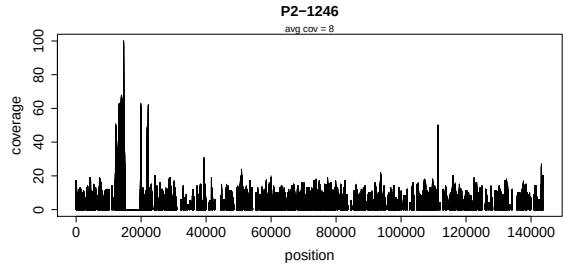
TABLE C.1: Processing of IM samples.



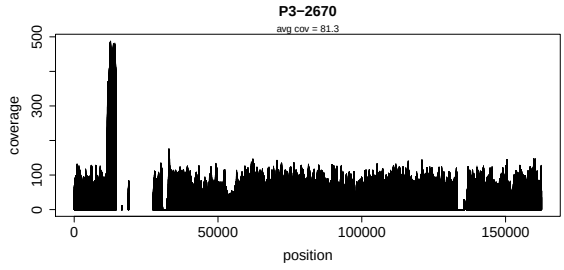
(A) P1-812



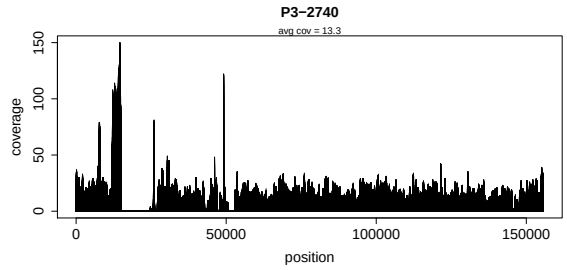
(B) P2-1213



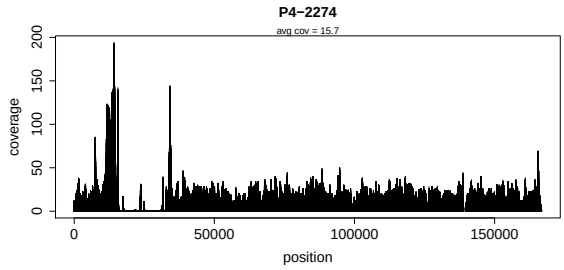
(C) P2-1246



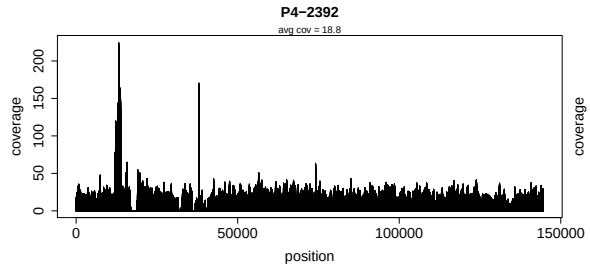
(D) P3-2670



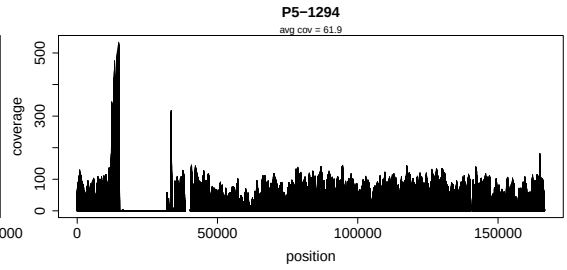
(E) P3-2740



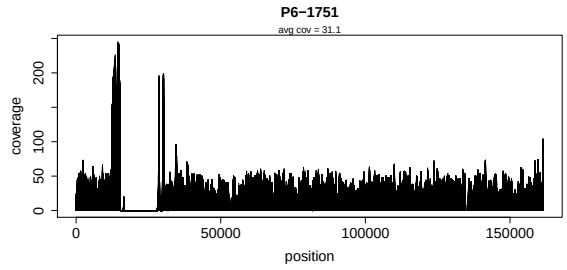
(F) P4-2274



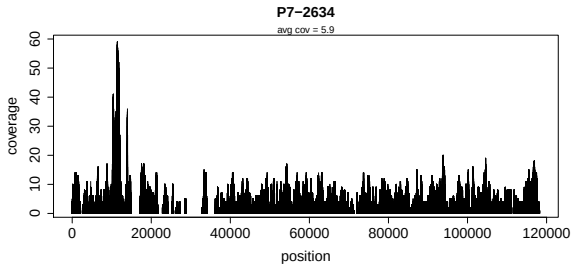
(G) P3-2392



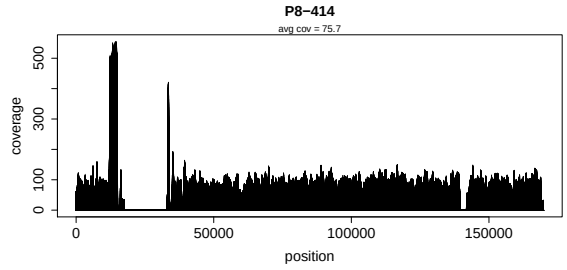
(H) P5-1294



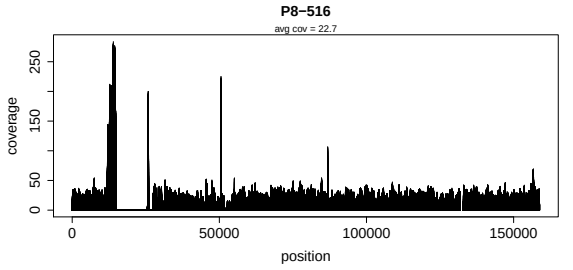
(I) P6-1751



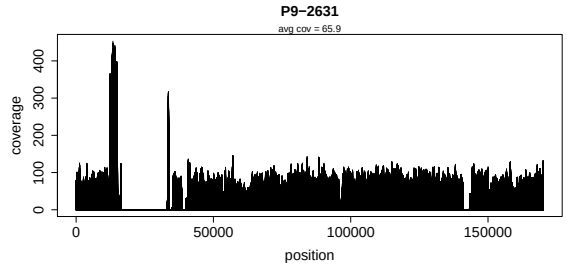
(J) P7-2634



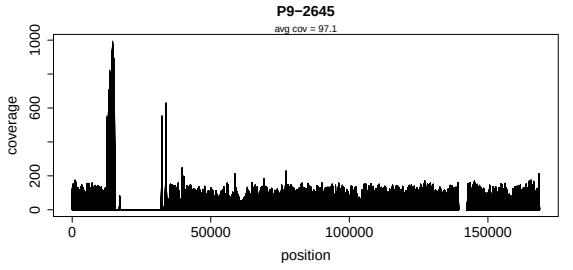
(K) P8-414



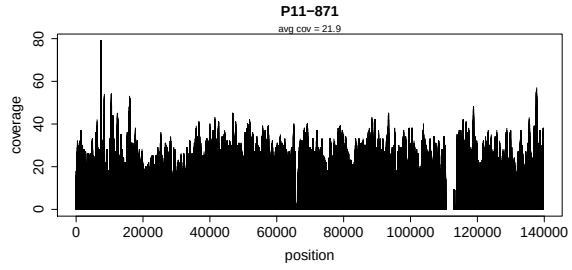
(L) P8-516



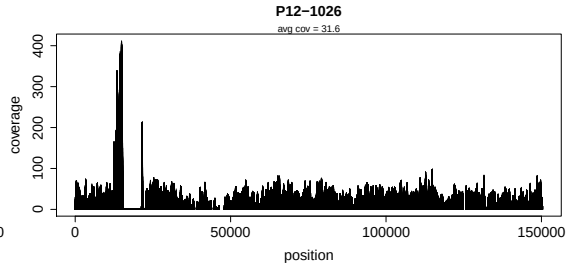
(M) P9-2631



(N) P9-2645



(O) P11-871



(P) P12-1026

FIGURE C.3: Coverage plots of IM samples after duplicate removal. Mapping is done against the sample consensus sequence directly after assembly (i.e. repeat regions are not masked or considered specifically).

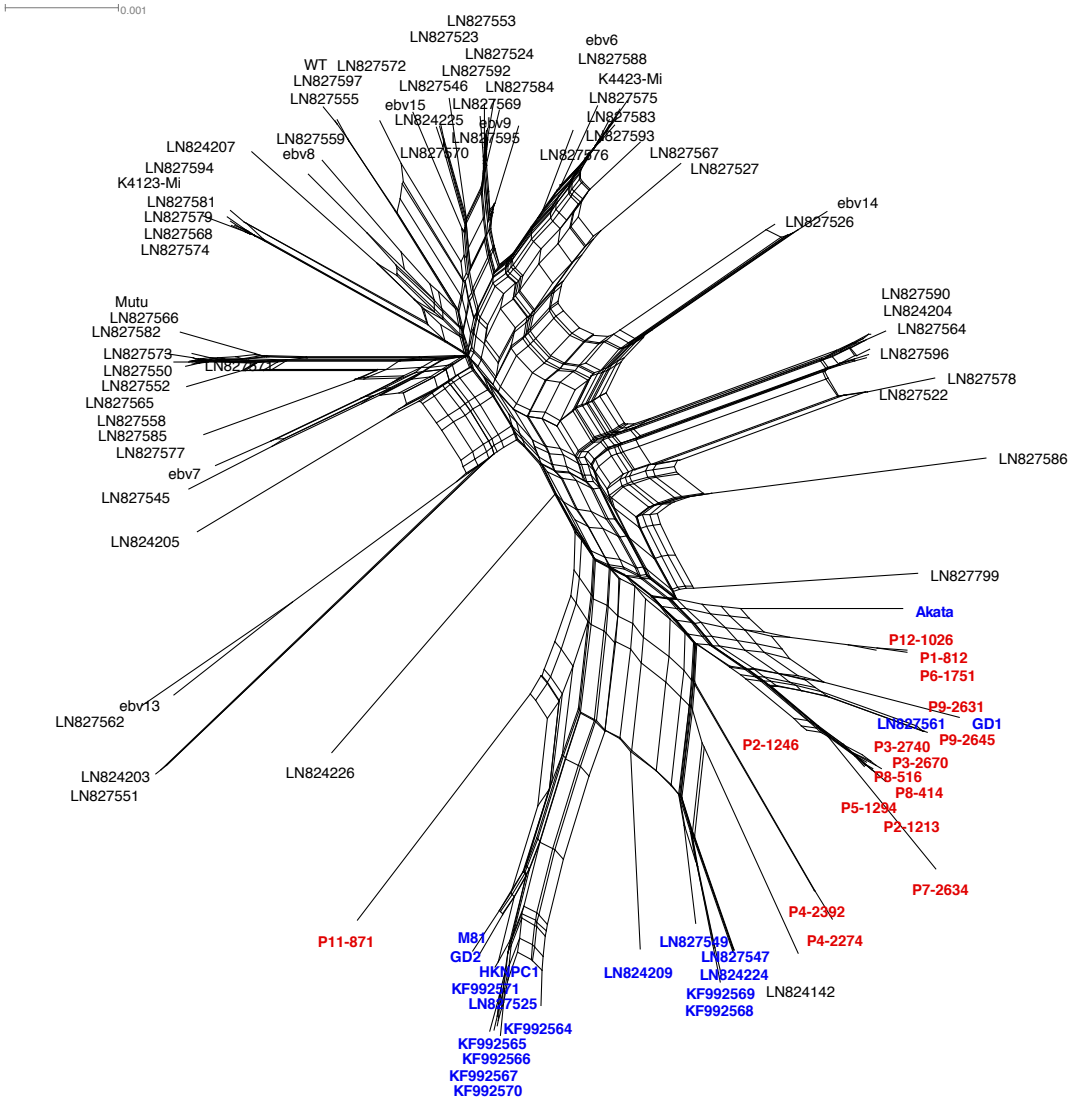


FIGURE C.4: Split network of whole genome alignment including the Japanese IM samples from chapter 5 as further representatives of Asian non-NPC genomes. The IM isolates from Japan are coloured in red, the remaining Asian isolates are coloured in blue.

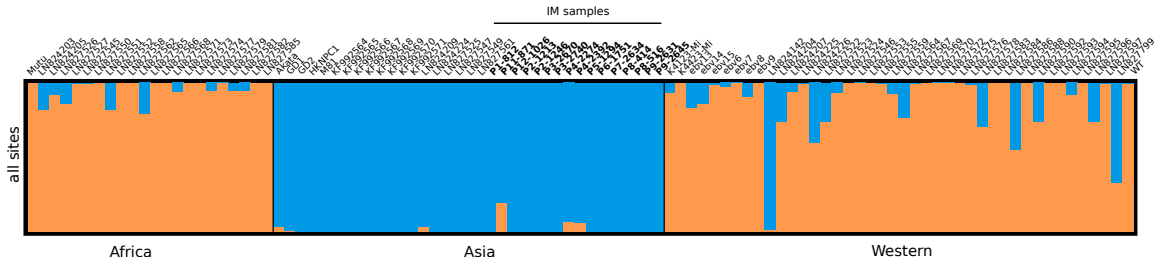


FIGURE C.5: Structure analysis of EBV type 1 genomes including the Japanese IM samples. All Japanese samples were completely or to a large proportion assigned to the blue (Asian) cluster.

	Pos	Ref	Var	Cov	Read1	Read2	Freq	ORF	nonsyn
P1-812	37856	T	C	28	22	6	21.43		
	80710	C	G	20	16	4	20		
	81070	C	T	30	22	8	26.67		
	81110	G	A	34	26	8	23.53	EBNA3A	A357T
	84503	C	A	20	16	4	20		
	99057	C	T	20	14	6	30		
	107643	A	C	20	14	6	30	BBRF3	M266L
	114682	T	G	20	16	4	20		
	140600	T	G	21	15	6	28.57	LF3	K915T
	140688	G	A	23	17	6	26.09	LF3	H886Y
	149989	G	A	30	24	6	20	LF2	
	152710	T	C	20	16	4	20		
	152854	T	C	20	16	4	20		
	162132	A	G	26	18	8	30.77		
	165354	T	C	20	16	4	20		
P2-1213	48433	C	T	20	16	4	20		
	60525	C	A	20	16	4	20	BOLF1	R809M
	60860	A	C	28	20	8	28.57	BOLF1	D697E
	60872	A	G	26	20	6	23.08		
	71085	C	A	20	16	4	20	BMLF1	R229M
								BSLF2/BMLF1	R270M
	86576	A	G	34	26	8	23.53	EBNA3C	I141V
P4-2274	9593	C	T	20	16	4	20		
	79739	G	A	20	16	4	20		
	81722	A	T	20	16	4	20		
	86232	T	G	35	26	9	25.71	EBNA3C	Y51D
	90389	G	T	20	15	5	25		
	90412	A	G	23	18	5	21.74		
	91046	T	C	20	16	4	20		
	108600	T	A	20	16	4	20	BBLF1	Y29F
	116286	C	A	20	16	4	20		
	117231	C	T	34	26	8	23.53		
	117291	A	T	28	22	6	21.43		
	117303	A	G	28	22	6	21.43		
	118210	G	C	32	24	8	25		
	133947	C	T	20	16	4	20		
	134064	T	C	20	12	8	40		
	140688	G	A	26	20	6	23.08		
	145934	A	G	20	16	4	20		
	146055	A	G	24	16	8	33.33		
	146075	G	A	26	16	10	38.46		
	146110	T	C	28	16	12	42.86		
	146273	A	G	22	14	8	36.36		
	146706	T	C	20	16	4	20		
	146760	C	G	20	16	4	20		
	146772	A	G	20	16	4	20		
	146780	T	C	20	16	4	20		
	148623	A	G	24	18	6	25		
	148652	T	G	24	16	8	33.33		
	149146	T	G	20	16	4	20	LF2	

	Pos	Ref	Var	Cov	Read1	Read2	Freq	ORF	nonsyn
	157839	T	C	20	16	4	20		
	168346	T	G	20	16	4	20		
	168845	T	G	24	16	8	33.33		
P5-1294	96720	C	T	20	16	4	20	EBNA1	R354W
P6-1751	37856	T	C	32	24	8	25		
	63911	C	T	20	16	4	20		
	67621	A	T	26	20	6	23.08		
	78005	T	A	26	20	6	23.08		
	79649	C	T	26	20	6	23.08		
	81984	C	G	25	20	5	20		
	84257	A	G	30	24	6	20		
	84524	A	T	30	24	6	20	EBNA3B	Q461H
	86779	T	C	20	16	4	20		
	87256	G	A	30	22	8	26.67		
	87265	C	T	28	22	6	21.43		
	89920	G	A	20	16	4	20		
	99285	A	C	20	12	8	40		
	106006	A	G	21	14	7	33.33		
	118210	C	G	28	22	6	21.43	BDLF3	V190L
	123264	A	C	20	16	4	20		
	140001	T	G	40	32	8	20		
	147747	G	C	28	20	8	28.57		
	149017	T	A	26	18	8	30.77		
	149326	C	T	20	12	8	40	LF2	
	149370	T	C	22	14	8	36.36	LF2	
P9-2631	88491	C	A	80	52	28	35		
P11-871	149055	A	G	35	27	8	22.86	LF2	
P12-1026	80349	G	C	46	36	10	21.74		
	81070	C	T	50	40	10	20		
	81212	T	C	38	30	8	21.05		
	81682	A	G	32	24	8	25		
	81722	T	A	24	18	6	25	EBNA3A	F561I
	82201	C	A	38	30	8	21.05		
	82239	G	A	44	32	12	27.27	EBNA3A	R733Q
	83092	A	G	36	28	8	22.22	EBNA3B	Q10R
	83186	G	A	34	22	12	35.29		
	83192	A	G	34	22	12	35.29		
	83237	A	G	30	18	12	40		
	83283	G	C	37	24	13	35.14	EBNA3B	E74Q
	84653	G	A	20	16	4	20		
	86779	T	C	26	20	6	23.08		
	88752	T	C	28	22	6	21.43	EBNA3C	L866S
	89081	A	G	42	30	12	28.57	EBNA3C	K976E
	89920	A	G	30	14	16	46.67		
	90046	A	C	46	32	14	30.43	BZLF1	S205A
	90107	A	G	52	40	12	23.08		
	91005	C	A	44	34	10	22.73		
	91011	C	T	40	28	12	30		
	93087	A	G	41	31	10	24.39		
	93194	A	C	64	48	16	25	BRRF1	D101A

Pos	Ref	Var	Cov	Read1	Read2	Freq	ORF	nonsyn
93225	A	G	64	48	16	25		
93294	T	C	44	34	10	22.73		
94840	A	G	30	22	8	26.67		
94900	G	T	20	12	8	40		
94950	G	A	20	12	8	40	BRRF2	R313H
97848	C	T	42	32	10	23.81		
97951	A	C	26	16	10	38.46		
97972	G	A	26	16	10	38.46	BKRF2	G102S
98005	A	T	20	14	6	30	BKRF2	T113S
102988	G	A	22	16	6	27.27		
104549	C	A	20	16	4	20	BBLF2-BBLF3	M705I
104635	G	A	20	16	4	20	BBLF2-BBLF3	H677Y
108561	A	G	26	20	6	23.08	BBLF1	V42A
112914	T	G	38	28	10	26.32	BGRF1-BDRF1	I89S
116501	C	A	20	16	4	20	BDLF4	D78Y
116664	G	A	20	16	4	20		
117030	C	T	28	16	12	42.86		
117093	C	G	22	12	10	45.45		
117231	T	C	30	24	6	20		
117291	T	A	34	26	8	23.53		
117303	G	A	40	30	10	25		
117360	A	G	26	20	6	23.08		
118202	G	T	30	24	6	20		
118372	C	A	20	16	4	20	BDLF3	A136S
130603	T	C	33	22	11	33.33		
148447	C	T	31	22	9	29.03		
149146	G	T	28	22	6	21.43	LF2	
149326	C	T	46	36	10	21.74	LF2	
152260	G	T	34	24	10	29.41		
164292	T	G	25	20	5	20		
164633	T	G	20	12	8	40		
166230	C	A	30	22	8	26.67	LMP2	S38Y
166353	C	A	20	16	4	20	LMP2	T79N
166362	A	C	20	14	6	30	LMP2	Q82P
166367	C	T	24	16	8	33.33		
167172	C	T	40	32	8	20		
169090	C	A	28	20	8	28.57		

TABLE C.2: Minority variants of IM samples. Pos: WT position. Ref: Consensus base. Var: Minor variant base. Cov: Read depth at this position. Read1/2: Number of reads supporting the consensus or variant base, respectively. Freq: Minor variant frequency. ORF: Open reading frame. nonsyn: Amino acid change from consensus to variant at the respective protein position.

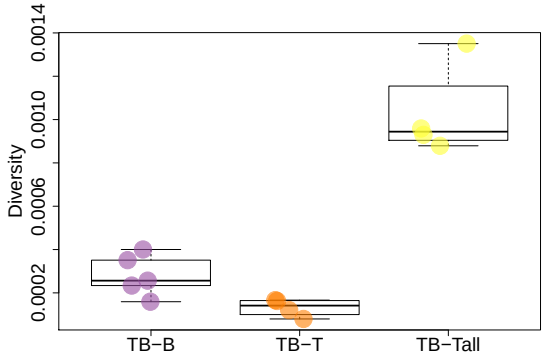


FIGURE C.6: Intrahost diversity for the paired tumour and blood samples. The line in the box indicates the median, the bottom and top of the box show the 25th and 75th percentile. TB-B: blood samples, TB-T: subsampled tumour samples, TB-Tall: complete tumour sample.

ORF	pos WT	pos aa	Epitope ID	Peptide
LMP2	207	169	20892	GLGTLGAA**L**
LMP2	410	208	60930	**S**SCSSCPLSK
LMP2	410	208	60931	**S**SCSSCPLSKI
LMP2	613	248	105557	TVCGGI**M**FL
LMP2	1068	348	6808	CPL**S**KILL
LMP2	1075	350	6808	CPLSK**I**LL
EBNA3A	79975	8	5679	AW**N**AGFLRGRAYGLD
EBNA3A	79976	8	5679	AW**N**AGFLRGRAYGLD
EBNA3A	80609	190	63546	TETAQAW**N**AGFLRGRAYGIDLLRTE
EBNA3A	80626	195	63546	TETAQAWNAGFL**R**GRAYGIDLLRTE
EBNA3A	81416	459	75356	Y**P**LHEQHGM
EBNA3A	81432	464	75356	YPLHEQ**H**GM
EBNA3B	84345	402	5316	AVF**D**RKSDAK
EBNA3B	84413	424	27375	ILTDFSVI**K**
EBNA3B	84413	424	29466	IVTDFSVI**K**
EBNA3B	85116	659	68229	VE**I**TPYKPTW
EBNA3C	86391	104	47807	PHDI**T**PYTARNIRDAACRAV
EBNA3C	86400	107	47807	PHDITPY**T**ARNIRDAACRAV
EBNA3C	86576	141	26981	**I**LCFVMAARQRLQDI
EBNA3C	87163	336	30430	K**E**HVIQNAF
BZLF1	90158	195	53128	RAKFKQ**L**L
BZLF1	90158	195	53129	RAKFKQ**L**LQ
BZLF1	90379	163	23103	GVPQPAPVAAPARRTRKP**Q**QPE
BZLF1	90412	152	23103	GVPQPAP**V**AAPARRTRKPQQPE
BZLF1	90430	146	23103	G**V**PQPAPVAAPARRTRKPQQPE
BZLF1	90665	68	149830	LTAYHVS**T**APTGSWF
BZLF1	90684	61	13701	EPLPQGQ**L**TAY
BZLF1	90684	61	38458	LPEPLPQGQ**L**TAY
BZLF1	90684	61	149830	**L**TAYHVSTAPTGSWF

TABLE C.3: Experimentally described epitopes that are affected by the nonsynonymous SNPs found to differentiate between Asian NPC and Non-NPC genomes. ORF: open reading frame. pos WT: position in WT genome. pos aa: position in protein. Epitope ID: epitope identifier from IEDB database. Peptide: epitope sequence in IEDB, the affected site is bold.

Bibliography

Abdirad, A et al. (2007). "Epstein-Barr virus associated gastric carcinoma: a report from Iran in the last four decades". In: *Diagn Pathol* 2, p. 25. DOI: 1746-1596-2-25[pii]\r10.1186/1746-1596-2-25[doi].

Ablashi, Dharam V. and John M. Easton (1976). "Preventive Vaccination against Herpesvirus saimiri-induced Neoplasia". In: *Cancer Research* 36, pp. 701–703. ISSN: 15387445.

Adams, Alice (1987). "Replication of Latent Epstein-Barr Virus Genomes in Raji Cells". In: *Journal of Virology* 61.5, pp. 1743–1746. ISSN: 0022-538X.

Ai, Junhong et al. (2012). "Analysis of EBNA-1 and LMP-1 variants in diseases associated with EBV infection in Chinese children." In: *Virology journal* 9, p. 13. ISSN: 1743-422X. DOI: 10.1186/1743-422X-9-13.

Al-Mozaini, Maha et al. (2009). "Epstein-Barr virus BART gene expression". In: *Journal of General Virology* 90.2, pp. 307–316. ISSN: 00221317. DOI: 10.1099/vir.0.006551-0.

Alfieri, C et al. (1996). "Epstein-Barr virus transmission from a blood donor to an organ transplant recipient with recovery of the same virus strain from the recipient's blood and oropharynx." In: *Blood* 87.2, pp. 812–7. ISSN: 0006-4971. DOI: 10.1016/S0887-7963(96)80066-6.

Allen, Michael D, Lawrence S Young, and Christopher W Dawson (2005). "The Epstein-Barr virus-encoded LMP2A and LMP2B proteins promote epithelial cell spreading and motility." In: *Journal of virology* 79.3, pp. 1789–802. ISSN: 0022-538X. DOI: 10.1128/JVI.79.3.1789-1802.2005.

Altschul, S F et al. (1990). "Basic local alignment search tool." In: *Journal of molecular biology* 215.3, pp. 403–10. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.

Ambinder, Richard F. and Risa B. Mann (1994). "Epstein-Barr-encoded RNA in situ hybridization: Diagnostic applications". In: *Human Pathology* 25.6, pp. 602–605. ISSN: 00468177. DOI: 10.1016/0046-8177(94)90227-5.

Anderson, Leah J. and Richard Longnecker (2008). "EBV LMP2A provides a surrogate pre-B cell receptor signal through constitutive activation of the ERK/MAPK pathway". In: *Journal of General Virology* 89.7, pp. 1563–1568. ISSN: 00221317. DOI: 10.1099/vir.0.2008/001461-0.

Anisimova, Maria, Joseph P Bielawski, and Ziheng Yang (2002). "Accuracy and power of bayes prediction of amino acid sites under positive selection." In: *Molecular biology and evolution* 19, pp. 950–958. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a004152.

Anisimova, Maria, Rasmus Nielsen, and Ziheng Yang (2003). "Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid

sites". In: *Genetics* 164.3, pp. 1229–1236. ISSN: 00166731. DOI: 10.1093/bioinformatics/btn086.

Anvret, Maria, Anne Karlsson, and Gunnar Bjursell (1984). "Evidence for integrated EBV genomes in raji cellular DNA". In: *Nucleic Acids Research* 12.2, pp. 1149–1161. ISSN: 03051048. DOI: 10.1093/nar/12.2.1149.

Ariza, Maria-Eugenia et al. (2009). "The EBV-encoded dUTPase activates NF-kappa B through the TLR2 and MyD88-dependent signaling pathway." In: *Journal of immunology (Baltimore, Md. : 1950)* 182, pp. 851–859. ISSN: 1550-6606. DOI: 10.4049/jimmunol.182.2.851.

Arriola, Edurne et al. (2007). "Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation". In: *Laboratory Investigation* 87.1, pp. 75–83. ISSN: 0023-6837. DOI: 10.1038/labinvest.3700495.

Aubry, V. et al. (2014). "Epstein-Barr Virus Late Gene Transcription Depends on the Assembly of a Virus-Specific Preinitiation Complex". In: *Journal of Virology* 88.21, pp. 12825–12838. ISSN: 0022-538X. DOI: 10.1128/JVI.02139-14.

Ayadi, Wajdi et al. (2007). "Polymorphism analysis of Epstein-Barr virus isolates of nasopharyngeal carcinoma biopsies from Tunisian patients". In: *Virus Genes* 34.2, pp. 137–145. ISSN: 09208569. DOI: 10.1007/s11262-006-0051-2.

Babcock, G J et al. (1998). "EBV persistence in memory B cells in vivo." In: *Immunity* 9.3, pp. 395–404. ISSN: 1074-7613. DOI: S1074-7613(00)80622-6[pii].

Babcock, Gregory J, Donna Hochberg, and David A Thorley-Lawson (2000). "The Expression Pattern of Epstein-Barr Virus Latent Genes In Vivo Is Dependent upon the Differentiation Stage of the Infected B Cell". In: *Immunity* 13.4, pp. 497–506. ISSN: 10747613. DOI: 10.1016/S1074-7613(00)00049-2.

Baer, R et al. (1984). "DNA sequence and expression of the B95–8 Epstein-Barr virus genome". In: *Nature* 310, pp. 207–211.

"Handbook of Statistical Genetics" (2007). In: ed. by D J Balding, M Bishop, and C Cannings. 3rd. Chap. 27: Linkage Disequilibrium, Recombination and Selection.

Balfour, Henry H. et al. (2013). "Behavioral, virologic, and immunologic factors associated with acquisition and severity of primary epstein-barr virus infection in university students". In: *Journal of Infectious Diseases* 207.1, pp. 80–88. ISSN: 00221899. DOI: 10.1093/infdis/jis646.

Balfour, Henry H Jr, Samantha K Dunmire, and Kristin A Hogquist (2015). "Infectious mononucleosis." In: *Clinical & translational immunology* 4.2, e33. ISSN: 2050-0068 (Electronic). DOI: 10.1038/cti.2015.1.

Bankevich, Anton et al. (2012). "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing". In: *Journal of Computational Biology* 19.5, pp. 455–477. ISSN: 1066-5277. DOI: 10.1089/cmb.2012.0021.

Banko, Ana V. et al. (2016). "Characterization of the variability of Epstein-Barr virus genes in nasopharyngeal biopsies: Potential predictors for carcinoma progression". In: *PLoS ONE* 11.4. ISSN: 19326203. DOI: 10.1371/journal.pone.0153498.

Barth, Stephanie et al. (2008). "Epstein-Barr virus-encoded microRNA miR-BART2 down-regulates the viral DNA polymerase BALF5". In: *Nucleic Acids Research* 36.2, pp. 666–675. ISSN: 03051048. DOI: 10.1093/nar/gkm1080.

Basson, V and A A Sharp (1969). "Monospot: a differential slide test for infectious mononucleosis." In: *Journal of clinical pathology* 22.3, pp. 324–5. ISSN: 0021-9746. DOI: 10.1136/jcp.22.3.324.

Bechtel, D et al. (2005). "Transformation of BCR-deficient germinal center B cells by EBV supports a major role of the virus in the pathogenesis of Hodgkin and post transplant lymphoma". In: *Blood* 106.13, pp. 4345–4350. DOI: 10.1182/blood-2005-06-2342.Supported.

Bei, Jin-Xin et al. (2016). "Genetic susceptibility to the endemic form of NPC". In: *Chinese Clinical Oncology* 5.2. ISSN: 2304-3873.

Bellan, Cristiana et al. (2005). "Immunoglobulin gene analysis reveals 2 distinct cells of origin for EBV-positive and EBV-negative Burkitt lymphomas". In: *Blood* 106.3, pp. 1031–1036. ISSN: 00064971. DOI: 10.1182/blood-2005-01-0168.

BenAyed-Guerfali, Dorra et al. (2011). "Characteristics of epstein barr virus variants associated with gastric carcinoma in Southern Tunisia." In: *Virology journal* 8.1, p. 500. ISSN: 1743-422X. DOI: 10.1186/1743-422X-8-500.

Berger, C et al. (1997). "The 30-bp deletion variant of Epstein-Barr virus-encoded latent membrane protein-1 prevails in acute infectious mononucleosis." In: *The Journal of infectious diseases* 176.5, pp. 1370–1373. ISSN: 0022-1899 (Print).

Berger, C et al. (2001). "Dynamics of Epstein-Barr virus DNA levels in serum during EBV-associated disease." In: *J. Med. Virol.* 64.4, pp. 505–12. ISSN: 0146-6615.

Berger, Christoph et al. (1999). "Sequence polymorphisms between latent membrane proteins LMP1 and LMP2A do not correlate in EBV-associated reactive and malignant lympho- proliferations". In: *International Journal of Cancer* 81.3, pp. 371–375. ISSN: 00207136. DOI: 10.1002/(SICI)1097-0215(19990505)81:3<371::AID-IJC10>3.0.CO;2-D.

Bhatia, K et al. (1996). "Variation in the sequence of Epstein Barr virus nuclear antigen 1 in normal peripheral blood lymphocytes and in Burkitt's lymphomas." In: *Oncogene* 13.1, pp. 177–181. ISSN: 0950-9232 (Print).

Blaes, Anne H. et al. (2005). "Rituximab therapy is effective for posttransplant lymphoproliferative disorders after solid organ transplantation: Results of a phase II trial". In: *Cancer* 104.8, pp. 1661–1667. ISSN: 0008543X. DOI: 10.1002/cncr.21391.

Blake, N et al. (1999). "Inhibition of antigen presentation by the glycine/alanine repeat domain is not conserved in simian homologues of Epstein-Barr virus nuclear antigen 1". In: *Journal of virology* 73.9, pp. 7381–7389. ISSN: 0022-538X.

Bollard, Catherine M. et al. (2014). "Sustained complete responses in patients with lymphoma receiving autologous cytotoxic T lymphocytes targeting Epstein-Barr virus latent membrane proteins". In: *Journal of Clinical Oncology* 32.8, pp. 798–808. ISSN: 15277755. DOI: 10.1200/JCO.2013.51.5304.

Borza, Corina M and Lindsey M Hutt-Fletcher (2002). "Alternate replication in B cells and epithelial cells switches tropism of Epstein-Barr virus." In: *Nature medicine* 8.6, pp. 594–599. ISSN: 10788956. DOI: 10.1038/nm0602-594.

Bouzid, M. et al. (1994). "Epstein-Barr virus genotypes in NPC biopsies from North Africa". In: *International Journal of Cancer* 56.4, pp. 468–473. ISSN: 10970215. DOI: 10.1002/ijc.2910560403.

Bräuninger, Andreas et al. (2006). *Molecular biology of Hodgkin's and Reed/Sternberg cells in Hodgkin's lymphoma*. DOI: 10.1002/ijc.21716.

Bredel, Markus et al. (2005). "Amplification of whole tumor genomes and gene-by-gene mapping of genomic aberrations from limited sources of fresh-frozen and paraffin-embedded DNA." In: *The Journal of molecular diagnostics : JMD* 7.2, pp. 171–182. ISSN: 15251578. DOI: 10.1016/S1525-1578(10)60543-0.

Brown, Amanda C. et al. (2015). "Rapid whole-genome sequencing of mycobacterium tuberculosis isolates directly from clinical samples". In: *Journal of Clinical Microbiology* 53.7, pp. 2230–2237. ISSN: 1098660X. DOI: 10.1128/JCM.00486-15.

Brown, Jay C. (2014). "The role of DNA repair in herpesvirus pathogenesis". In: *Genomics* 104.4, pp. 287–294. ISSN: 10898646. DOI: 10.1016/j.ygeno.2014.08.005.

Bruen, Trevor C, Hervé Philippe, and David Bryant (2006). "A simple and robust statistical test for detecting the presence of recombination." In: *Genetics* 172.4, pp. 2665–81. ISSN: 0016-6731. DOI: 10.1534/genetics.105.048975.

Buell, Philip (1974). "The Effect of Migration on the Risk of Nasopharyngeal Cancer among Chinese". In: *Cancer Research* 34.5, pp. 1189–1191. ISSN: 15387445.

Burkitt, Denis (1958). "A sarcoma involving the jaws in african children". In: *British Journal of Surgery* 46.197, pp. 218–223. ISSN: 13652168. DOI: 10.1002/bjs.18004619704.

Bushnell, B. *BBMap*. http://sourceforge.net/projects/bbmap.

Cai, Xuezhong et al. (2006). "Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed". In: *PLoS Pathogens* 2.3, pp. 0236–0247. ISSN: 15537366. DOI: 10.1371/journal.ppat.0020023.

Calderwood, Michael A et al. (2007). "Epstein-Barr virus and virus human protein interaction maps." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.18, pp. 7606–11. ISSN: 0027-8424. DOI: 10.1073/pnas.0702332104.

Caldwell, Robert G. et al. (1998). "Epstein-Barr virus LMP2A drives B cell development and survival in the absence of normal B cell receptor signals". In: *Immunity* 9.3, pp. 405–411. ISSN: 10747613. DOI: 10.1016/S1074-7613(00)80623-8.

Callan, M F et al. (1998). "Direct visualization of antigen-specific CD8+ T cells during the primary immune response to Epstein-Barr virus In vivo". In: *J Exp Med* 187.9, pp. 1395–1402.

Canaan, Allon et al. (2009). "EBNA1 regulates cellular gene expression by binding cellular promoters." In: *Proceedings of the National Academy of Sciences of the United States of America* 106.52, pp. 22421–22426. ISSN: 0027-8424. DOI: 10.1073/pnas.0911676106.

Cannon, J S et al. (2000). "A new primary effusion lymphoma-derived cell line yields a highly infectious Kaposi's sarcoma herpesvirus-containing supernatant." In: *Journal*

of virology 74.21, pp. 10187–10193. ISSN: 0022-538X. DOI: 10.1128/JVI.74.21.10187-10193.2000.

Capello, Daniela, Davide Rossi, and Gianluca Gaidano (2005). "Post-transplant lymphoproliferative disorders: molecular basis of disease histogenesis and pathogenesis." In: *Hematological oncology* 23.2, pp. 61–7. ISSN: 0278-0232. DOI: 10.1002/hon.751.

Cepok, Sabine et al. (2005). "Identification of Epstein-Barr virus proteins as putative targets of the immune response in multiple sclerosis". In: *The Journal of Clinical Investigation* 115.5, pp. 1352–1360. ISSN: 0021-9738. DOI: 10.1172/JCI23661.

Chaganti, Sridhar et al. (2005). "Epstein-Barr virus infection in vitro can rescue germinal center B cells with inactivated immunoglobulin genes". In: *Blood* 106.13, pp. 4249–4252. ISSN: 00064971. DOI: 10.1182/blood-2005-06-2327.

Chang, Cindy M et al. (2009). "The extent of genetic diversity of Epstein-Barr virus and its geographic and disease patterns: a need for reappraisal." In: *Virus research* 143.2, pp. 209–21. ISSN: 1872-7492. DOI: 10.1016/j.virusres.2009.07.005.

Chao, Mei et al. (2015). "The V-val subtype Epstein-Barr virus nuclear antigen 1 promotes cell survival after serum withdrawal". In: *Oncology reports* 33 (2), pp. 958–966.

Chen, C. J. et al. (1990). "Multiple risk factors of nasopharyngeal carcinoma: Epstein-Barr virus, malarial infection, cigarette smoking and familial tendency". In: *Anticancer Research* 10.2 B, pp. 547–553. ISSN: 02507005.

Chen, Jian ning et al. (2011). "Epstein-Barr virus genome polymorphisms of Epstein-Barr virus-associated gastric carcinoma in gastric remnant carcinoma in Guangzhou, southern China, an endemic area of nasopharyngeal carcinoma". In: *Virus Research* 160.1-2, pp. 191–199. ISSN: 01681702. DOI: 10.1016/j.virusres.2011.06.011.

Chen, Jian Ning et al. (2010a). "Association of distinctive Epstein-Barr virus variants with gastric carcinoma in Guangzhou, Southern China". In: *Journal of Medical Virology* 82.4, pp. 658–667. ISSN: 01466615. DOI: 10.1002/jmv.21731.

Chen, Shu Jen et al. (2010b). "Characterization of epstein-barr virus miRNAome in nasopharyngeal carcinoma by deep sequencing". In: *PLoS ONE* 5.9, pp. 1–14. ISSN: 19326203. DOI: 10.1371/journal.pone.0012745.

Chen, T. Scott, Aaron W. Reinke, and Amy E. Keating (2011). "Design of peptide inhibitors that bind the bZIP domain of Epstein-Barr virus protein BZLF1". In: *Journal of Molecular Biology* 408.2, pp. 304–320. ISSN: 00222836. DOI: 10.1016/j.jmb.2011.02.046.

Chêne, Arnaud et al. (2007). "A Molecular Link between Malaria and Epstein–Barr Virus Reactivation". In: *PLoS Pathog* 3.6, e80. DOI: 10.1371/journal.ppat.0030080.

Chesnokova, Liudmila S and Lindsey M Hutt-Fletcher (2011). "Fusion of EBV with epithelial cells can be triggered by {alpha}v{beta}5 in addition to {alpha}v{beta}6 and {alpha}v{beta}8 and integrin binding triggers a conformational change in gHgL." In: *Journal of virology* September. ISSN: 1098-5514. DOI: 10.1128/JVI.05580-11.

Chhatre, Vikram E and Kevin J Emerson (2016). "StrAuto: Automation and parallelization of STRUCTURE analysis". In: *http://strauto.popgen.org*.

Chiara, Matteo et al. (2016). "Geographic population structure in Epstein-Barr Virus revealed by comparative genomics". In: *Genome Biology and Evolution*. DOI: 10.1093/gbe/evw226. eprint: http://gbe.oxfordjournals.org/content/early/2016/09/14/gbe.evw226.full.pdf+html.

Chiu, Christopher et al. (2014). "Broadly Reactive Human CD8 T Cells that Recognize an Epitope Conserved between VZV, HSV and EBV". In: *PLoS Pathogens* 10.3. ISSN: 15537374. DOI: 10.1371/journal.ppat.1004008.

Cho, Young-Gyu et al. (1999). "Evolution of Two Types of Rhesus Lymphocryptovirus Similar to Type 1 and Type 2 Epstein-Barr Virus". In: *Journal of Virology* 73.11, pp. 9206–9212. eprint: http://jvi.asm.org/content/73/11/9206.full.pdf+html.

Christiansen, M. T. et al. (2017). "Use of whole genome sequencing in the Dutch Acute HCV in HIV study: focus on transmitted antiviral resistance". In: *Clinical Microbiology and Infection* 23.2, 123.e1–123.e4. ISSN: 14690691. DOI: 10.1016/j.cmi.2016.09.018.

Christiansen, Mette T et al. (2014). "Whole-genome enrichment and sequencing of Chlamydia trachomatisdirectly from clinical samples". In: *BMC Infectious Diseases* 14.1, p. 591. ISSN: 1471-2334. DOI: 10.1186/s12879-014-0591-3.

Chua, Melvin L K et al. (2016). "Nasopharyngeal carcinoma". In: *The Lancet*. Vol. 387. 10022, pp. 1012–1024. ISBN: 0923-7534 (Print)\r0923-7534 (Linking). DOI: 10.1016/S0140-6736(15)00055-0.

Churchill, A. E., R. C. Chubb, and W. Baxendale (1969). "The attenuation, with loss of oncogenicity, of the herpes-type virus of Marek's disease (strain HPRS-16) on passage in cell culture." In: *Journal of General Virology* 4.4, pp. 557–564. ISSN: 00221317. DOI: 10.1099/0022-1317-4-4-557.

CLC Workbench 7. https://www.qiagenbioinformatics.com/.

Clute, S C et al. (2005). "Cross-reactive influenza virus-specific CD8+ T cells contribute to lymphoproliferation in Epstein-Barr virus-associated infectious mononucleosis". In: *The Journal of clinical investigation* 115.12, pp. 3602–3612. ISSN: 0021-9738. DOI: 10.1172/JCI25078.

Cohen, Jeffrey I (2015). "Epstein–barr virus vaccines". In: *Clinical & Translational Immunology* 4.1, e32. ISSN: 2050-0068. DOI: 10.1038/cti.2014.27.

Comoli, Patrizia et al. (2002). "Infusion of autologous Epstein-Barr virus (EBV)-specific cytotoxic T cells for prevention of EBV-related lymphoproliferative disorder in solid organ transplant recipients with evidence of active virus replication". In: *Blood* 99.7, pp. 2592–2598. ISSN: 00064971. DOI: 10.1182/blood.V99.7.2592.

Cooper, Andrew et al. (2003). "EBNA3A association with RBP-Jkappa down-regulates c-myc and Epstein-Barr virus-transformed lymphoblast growth." In: *Journal of virology* 77.2, pp. 999–1010. ISSN: 0022-538X. DOI: 10.1128/JVI.77.2.999-1010.2003.

Correa, Rita Mariel et al. (2004). "Epstein-Barr virus (EBV) in healthy carriers: Distribution of genotypes and 30 bp deletion in latent membrane protein-1 (LMP-1) oncogene". In: *Journal of Medical Virology* 73.4, pp. 583–588. ISSN: 01466615. DOI: 10.1002/jmv.20129.

Cosmopoulos, K et al. (2009). "Comprehensive profiling of Epstein-Barr virus microRNAs in nasopharyngeal carcinoma". In: *J Virol* 83.5, pp. 2357–2367. ISSN: 0022-538X. DOI: 10.1128/JVI.02104-08.

Csardi, Gabor and Tamas Nepusz (2006). "The igraph software package for complex network research". In: *InterJournal* Complex Systems, p. 1695.

Cui, Qian et al. (2016). "Nasopharyngeal carcinoma risk prediction via salivary detection of host and Epstein-Barr virus genetic variants". In: *Oncotarget* 5.0. ISSN: 1949-2553.

da Costa, Vivaldo G, Ariany C Marques-Silva, and Marcos L Moreli (2015). "The Epstein-Barr virus latent membrane protein-1 (LMP1) 30-bp deletion and XhoI-polymorphism in nasopharyngeal carcinoma: a meta-analysis of observational studies." In: *Systematic reviews* 4.1, p. 46. ISSN: 2046-4053. DOI: 10.1186/s13643-015-0037-z.

Dargan, Derrick J et al. (2010). "Sequential mutations associated with adaptation of human cytomegalovirus to growth in cell culture." In: *The Journal of general virology* 91.Pt 6, pp. 1535–46. ISSN: 1465-2099. DOI: 10.1099/vir.0.018994-0.

Davison, Andrew J (2007). "Chapter 2 Comparative analysis of the genomes". In: *Human Herpesviruses*. Ed. by Ann Arvin et al. Cambridge: Cambridge University Press. Chap. Comparative analysis of the genomes.

Davison, Andrew J. et al. (2009). *The order Herpesvirales*. DOI: 10.1007/s00705-008-0278-4.

Dawson, Christopher W., Rebecca J. Port, and Lawrence S. Young (2012). *The role of the EBV-encoded latent membrane proteins LMP1 and LMP2 in the pathogenesis of nasopharyngeal carcinoma (NPC)*. DOI: 10.1016/j.semcancer.2012.01.004.

de Jesus, O. (2003). "Updated Epstein-Barr virus (EBV) DNA sequence and analysis of a promoter for the BART (CST, BARF0) RNAs of EBV". In: *Journal of General Virology* 84.6, pp. 1443–1450. ISSN: 00221317. DOI: 10.1099/vir.0.19054-0.

Deane, Charlotte M. et al. (2002). "Protein Interactions Two Methods for Assessment of the Reliability of High Throughput Observations". In: *Molecular & Cellular Proteomics* 1.5, pp. 349–356. ISSN: 1535-9476, 1535-9484. DOI: 10.1074/mcp.M100037-MCP200.

Delecluse, H J et al. (1993). "Episomal and integrated copies of Epstein-Barr virus coexist in Burkitt lymphoma cell lines." In: *Journal of virology* 67.3, pp. 1292–1299. ISSN: 0022-538X.

Depledge, Daniel P et al. (2011). "Specific capture and whole-genome sequencing of viruses from clinical samples." In: *PloS one* 6.11, e27805. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027805.

Depledge, D.P. et al. (2014). "Deep sequencing of viral genomes provides insight into the evolution and pathogenesis of varicella zoster virus and its vaccine in humans". In: *Molecular Biology and Evolution* 31.2. ISSN: 07374038 15371719. DOI: 10.1093/molbev/mst210.

Dickson, R I and A D Flores (1985). "Nasopharyngeal carcinoma: an evaluation of 134 patients treated between 1971-1980." In: *The Laryngoscope* 95.3, pp. 276–83. ISSN: 0023-852X.

Do, Nguyen Van et al. (2008). "A major EBNA1 variant from Asian EBV isolates shows enhanced transcriptional activity compared to prototype B95.8". In: *Virus Research* 132.1-2, pp. 15–24. ISSN: 01681702. DOI: 10.1016/j.virusres.2007.10.020.

Dolan, Aidan et al. (2006). "The genome of Epstein-Barr virus type 2 strain AG876." In: *Virology* 350.1, pp. 164–70. ISSN: 0042-6822. DOI: 10.1016/j.virol.2006.01.015.

Doubrovina, Ekaterina et al. (2012). "Adoptive immunotherapy with unselected or EBV-specific T cells for biopsy-proven EBV + lymphomas after allogeneic hematopoietic cell transplantation". In: *Blood* 119.11, pp. 2644–2656. ISSN: 00064971. DOI: 10.1182/blood-2011-08-371971.

Duraiswamy, Jaikumar et al. (2003). "Ex vivo analysis of T-cell responses to Epstein-Barr virus-encoded oncogene latent membrane protein 1 reveals highly conserved epitope sequences in virus isolates from diverse geographic regions." In: *Journal of virology* 77.13, pp. 7401–10. ISSN: 0022-538X. DOI: 10.1128/JVI.77.13.7401.

Earl, Dent A. and Bridgett M. VonHoldt (2012). "STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method". In: *Conservation Genetics Resources* 4.2, pp. 359–361. ISSN: 18777252. DOI: 10.1007/s12686-011-9548-7.

Edwards, R H, F Seillier-Moiseiwitsch, and N Raab-Traub (1999). "Signature amino acid changes in latent membrane protein 1 distinguish Epstein-Barr virus strains." In: *Virology* 261.1, pp. 79–95. ISSN: 0042-6822. DOI: 10.1006/viro.1999.9855.

Edwards, Rachel Hood, Aron R Marquitz, and Nancy Raab-Traub (2008). "Epstein-Barr virus BART microRNAs are produced from a large intron prior to splicing." In: *Journal of virology* 82.18, pp. 9094–106. ISSN: 1098-5514. DOI: 10.1128/JVI.00785-08.

Ehlers, Bernhard et al. (2010). "Lymphocryptovirus phylogeny and the origins of Epstein-Barr virus". In: *Journal of General Virology* 91.3, pp. 630–642. ISSN: 00221317. DOI: 10.1099/vir.0.017251-0.

Elliott, Suzanne L et al. (2008). "Phase I trial of a CD8+ T-cell peptide epitope-based vaccine for infectious mononucleosis." In: *Journal of virology* 82.3, pp. 1448–57. ISSN: 1098-5514. DOI: 10.1128/JVI.01409-07.

Epeldegui, Marta et al. (2007). "Infection of human B cells with Epstein-Barr virus results in the expression of somatic hypermutation-inducing molecules and in the accrual of oncogene mutations". In: *Molecular Immunology* 44.5, pp. 934–942. ISSN: 01615890. DOI: 10.1016/j.molimm.2006.03.018.

Epstein, M A, B G Achong, and Y M Barr (1964). "Virus particles in cultured lymphoblasts from Burkitt's lymphoma." In: *The Lancet* 283.7335. Originally published as Volume 1, Issue 7335, pp. 702–703. ISSN: 0140-6736. DOI: http://dx.doi.org/10.1016/S0140-6736(64)91524-7.

Evanno, G., S. Regnaut, and J. Goudet (2005). "Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study". In: *Molecular Ecology* 14.8, pp. 2611–2620. ISSN: 09621083. DOI: 10.1111/j.1365-294X.2005.02553.x.

Fafi-Kremer, Samira et al. (2005). "Long-term shedding of infectious epstein-barr virus after infectious mononucleosis." In: *The Journal of infectious diseases* 191, pp. 985–989. ISSN: 0022-1899. DOI: 10.1086/428097.

Farina, Antonella et al. (2017). "Epstein-Barr virus lytic infection promotes activation of Toll-like receptor 8 innate immune response in systemic sclerosis monocytes". In: *Arthritis Research {&} Therapy* 19.1, p. 39. ISSN: 1478-6362. DOI: 10.1186/s13075-017-1237-9.

FASTX toolkit. http://hannonlab.cshl.edu/fastx_toolkit/index.html.

Feederle, Regina et al. (2011a). "A viral microRNA cluster strongly potentiates the transforming properties of a human herpesvirus". In: *PLoS Pathogens* 7.2. ISSN: 15537366. DOI: 10.1371/journal.ppat.1001294.

Feederle, Regina et al. (2011b). "The members of an Epstein-Barr virus microRNA cluster cooperate to transform B lymphocytes". In: *Journal of virology* 85.19, pp. 9801–9810. ISSN: 1098-5514. DOI: JVI.05100-11[pii]10.1128/JVI.05100-11.

Felsenstein, Joseph (1981). "Evolutionary trees from DNA sequences: A maximum likelihood approach". In: *Journal of Molecular Evolution* 17.6, pp. 368–376. ISSN: 00222844. DOI: 10.1007/BF01734359.

Feng, Fu-Tuo et al. (2015). "A single nucleotide polymorphism in the Epstein-Barr virus genome is strongly associated with a high risk of nasopharyngeal carcinoma". In: *Chinese Journal of Cancer* 34.3, p. 61. ISSN: 1944-446X. DOI: 10.1186/s40880-015-0073-z.

Ferlay, Jacques et al. (2015). "Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012". In: *International Journal of Cancer* 136.5, E359–E386. ISSN: 10970215. DOI: 10.1002/ijc.29210. arXiv: arXiv:1011.1669v3.

Fielding, C A et al. (2001). "Epstein-Barr virus LMP-1 natural sequence variants differ in their potential to activate cellular signaling pathways." In: *Journal of virology* 75.19, pp. 9129–9141. ISSN: 0022-538X (Print). DOI: 10.1128/JVI.75.19.9129-9141.2001.

Fixman, E D, G S Hayward, and S D Hayward (1992). "trans-acting requirements for replication of Epstein-Barr virus ori-Lyt." In: *Journal of virology* 66.8, pp. 5030–5039. ISSN: 0022-538X.

Fossum, Even et al. (2009). "Evolutionarily conserved herpesviral protein interaction networks". In: *PLoS Pathogens* 5.9. ISSN: 15537366. DOI: 10.1371/journal.ppat.1000570.

Fries, K L, W E Miller, and N Raab-Traub (1996). "Epstein-Barr virus latent membrane protein 1 blocks p53-mediated apoptosis through the induction of the A20 gene." In: *Journal of virology* 70.12, pp. 8653–8659. ISSN: 0022-538X.

Fu, Y X (2001). "Estimating mutation rate and generation time from longitudinal samples of DNA sequences." In: *Molecular biology and evolution* 18.4, pp. 620–6. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a003842.

Funk, Georg A, Rainer Gosert, and Hans H Hirsch (2007). "Viral dynamics in transplant patients : implications for disease". In: *Lancet Infectious Disease* 7.July, pp. 460–72.

Gahn, T a and B Sugden (1995). "An EBNA-1-dependent enhancer acts from a distance of 10 kilobase pairs to increase expression of the Epstein-Barr virus LMP gene." In: *Journal of virology* 69.4, pp. 2633–2636. ISSN: 0022-538X.

Gärtner, Barbara C. et al. (2002). "Evaluation of use of Epstein-Barr viral load in patients after allogeneic stem cell transplantation to diagnose and monitor posttransplant lymphoproliferative disease". In: *Journal of Clinical Microbiology* 40.2, pp. 351–358. ISSN: 00951137. DOI: 10.1128/JCM.40.2.351-358.2002.

Gerber, P et al. (1977). "Biologic and antigenic characteristics of Epstein-Barr virus-related Herpesviruses of chimpanzees and baboons". In: *International Journal of Cancer* 20 (3), pp. 448–59.

Gilligan, K et al. (1990). "Epstein-Barr virus small nuclear RNAs are not expressed in permissively infected cells in AIDS-associated leukoplakia." In: *Proceedings of the National Academy of Sciences of the United States of America* 87.22, pp. 8790–4. ISSN: 0027-8424.

Gires, Olivier et al. (1997). "Latent membrane protein 1 of Epstein-Barr virus mimics a constitutively active receptor molecule". In: *EMBO Journal* 16.20, pp. 6131–6140. ISSN: 02614189. DOI: 10.1093/emboj/16.20.6131.

Gonzalez-Farre, Blanca et al. (2014). "In vivo intratumoral Epstein-Barr virus replication is associated with XBP1 activation and early-onset post-transplant lymphoproliferative disorders with prognostic implications". In: *Modern Pathology* 27.12, pp. 1599–1611. ISSN: 0893-3952. DOI: 10.1038/modpathol.2014.68.

Goossens, T, U Klein, and R Küppers (1998). "Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease." In: *Proceedings of the National Academy of Sciences of the United States of America* 95.5, pp. 2463–2468. ISSN: 00278424. DOI: 10.1073/pnas.95.5.2463.

Gottschalk, S. (2001). "An Epstein-Barr virus deletion mutant associated with fatal lymphoproliferative disease unresponsive to therapy with virus-specific CTLs". In: *Blood* 97.4, pp. 835–843. ISSN: 00064971. DOI: 10.1182/blood.V97.4.835.

Gourzones, Claire et al. (2010). "Extra-cellular release and blood diffusion of BART viral micro-RNAs produced by EBV-infected nasopharyngeal carcinoma cells." In: *Virology journal* 7, p. 271. ISSN: 1743-422X. DOI: 10.1186/1743-422X-7-271.

Green, M and M G Michaels (2013). "Epstein-Barr virus infection and posttransplant lymphoproliferative disorder." In: *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 13 Suppl 3.September 2012, 41–54; quiz 54. ISSN: 1600-6143. DOI: 10.1111/ajt.12004.

Greenspan, J. S. et al. (1985). "Replication of Epstein-Barr virus within the epithelial cells of oral 'hairy' leukoplakia, an AIDS-associated lesion". In: *New England Journal of Medicine* 313.25, pp. 1564–1571. ISSN: 00284793. DOI: 10.1056/NEJM198512193132502.

Gruffat, H. et al. (2012). "The Epstein-Barr Virus BcRF1 Gene Product Is a TBP-Like Protein with an Essential Role in Late Gene Expression". In: *Journal of Virology* 86.11, pp. 6023–6032. ISSN: 0022-538X. DOI: 10.1128/JVI.00159-12.

Guiretti, Deisy M. et al. (2007). "Structural variability of the carboxy-terminus of Epstein-Barr virus encoded latent membrane protein 1 gene in Hodgkin's lymphomas." In: *Journal of medical virology* 79.11, pp. 1730–1722. ISSN: 01466615. DOI: 10.1002/jmv.21020.

Gulley, Margaret L and Weihua Tang (2010). "Using Epstein-Barr viral load assays to diagnose, monitor, and prevent posttransplant lymphoproliferative disorder." In: *Clinical microbiology reviews* 23.2, pp. 350–66. ISSN: 1098-6618. DOI: 10.1128/CMR.00006-09.

Gupta, S et al. (1996). "The maintenance of strain structure in populations of recombining infectious agents." In: *Nature medicine* 2.4, pp. 437–42. ISSN: 1078-8956. DOI: 10.1038/nm0496-437.

Gurevich, Alexey et al. (2013). "QUAST: Quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8, pp. 1072–1075. ISSN: 13674803. DOI: 10.1093/bioinformatics/btt086.

Gutiérrez, M. I. et al. (1997). "Sequence variations in EBNA-1 may dictate restriction of tissue distribution of Epstein-Barr virus in normal and tumour cells". In: *Journal of General Virology* 78.7, pp. 1663–1670. ISSN: 00221317. DOI: 10.1099/0022-1317-78-7-1663.

Gutiérrez, Marina I et al. (2002). "Discrete alterations in the BZLF1 promoter in tumor and non-tumor-associated Epstein-Barr virus." In: *Journal of the National Cancer Institute* 94.23, pp. 1757–63. ISSN: 0027-8874.

Görzer, Irene et al. (2006). "Characterization of Epstein-Barr virus Type I variants based on linked polymorphism among EBNA3A, -3B, and -3C genes". In: *Virus Research* 118.1–2, pp. 105 –114. ISSN: 0168-1702. DOI: https://doi.org/10.1016/j.virusres.2005.11.020.

Hadinoto, Vey et al. (2009). "The dynamics of EBV shedding implicate a central role for epithelial cells in amplifying viral output". In: *PLoS Pathogens* 5.7. ISSN: 15537366. DOI: 10.1371/journal.ppat.1000496.

Hage, Elias et al. (2017). "Characterization of Human Cytomegalovirus Genome Diversity in Immunocompromised Hosts by Whole-Genome Sequencing Directly From Clinical Specimens". In: *The Journal of Infectious Diseases* 215.11, p. 1673. DOI: 10.1093/infdis/jix157. eprint: /oup/backfile/content_public/journal/jid/215/11/10.1093_infdis_jix157/5/jix157.pdf.

Hammerschmidt, Wolfgang and Bill Sugden (1988). "Identification and characterization of oriLyt, a lytic origin of DNA replication of Epstein-Barr virus". In: *Cell* 55.3, pp. 427–433. ISSN: 00928674. DOI: 10.1016/0092-8674(88)90028-1.

Hammerschmidt, Wolfgang and Bill Sugden (2013). "Replication of Epstein-Barr viral DNA". In: *Cold Spring Harbor Perspectives in Biology* 5.1. ISSN: 19430264. DOI: 10.1101/cshperspect.a013029.

Han, Jing et al. (2012). "Sequence variations of latent membrane protein 2a in epstein-barr virus-associated gastric carcinomas from guangzhou, southern china". In: *PLoS ONE* 7.3. ISSN: 19326203. DOI: 10.1371/journal.pone.0034276.

Haque, T. et al. (2007). "Allogeneic cytotoxic T-cell therapy for EBV-positive posttransplantation lymphoproliferative disease: results of a phase 2 multicenter clinical trial P". In: *Blood* 110.4, pp. 1123–1132. DOI: 10.1182/blood-2006-12-063008.

Hasegawa, Masami, Hirohisa Kishino, and Taka aki Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". In: *Journal of Molecular Evolution* 22.2, pp. 160–174. ISSN: 00222844. DOI: 10.1007/BF02101694.

Haydon, Daniel T., Armanda D S Bastos, and Philip Awadalla (2004). "Low linkage disequilibrium indicative of recombination in foot-and-mouth disease virus gene sequence alignments". In: *Journal of General Virology* 85, pp. 1095–1100. ISSN: 00221317. DOI: 10.1099/vir.0.19588-0.

Heilmann, Andreas M F, Michael a Calderwood, and Eric Johannsen (2010). "Epstein-Barr virus LF2 protein regulates viral replication by altering Rta subcellular localization." In: *Journal of virology* 84.19, pp. 9920–31. ISSN: 1098-5514. DOI: 10.1128/JVI.00573-10.

Henle, G, W Henle, and C A Horwitz (1974). "Antibodies to Epstein-Barr virus associated nuclear antigen in infectious mononucleosis." In: *Journal of Infectious Diseases* 130, pp. 231–239.

Henle, W et al. (1987). "Antibody responses to Epstein-Barr virus-determined nuclear antigen (EBNA)-1 and EBNA-2 in acute and chronic Epstein-Barr virus infection." In: *Proceedings of the National Academy of Sciences of the United States of America* 84.2, pp. 570–4. ISSN: 0027-8424. DOI: 10.1073/pnas.84.2.570.

Heslop, Helen E. (2009). "How I treat EBV lymphoproliferation". In: *Blood* 114.19, pp. 4002–4008. ISSN: 00064971. DOI: 10.1182/blood-2009-07-143545.

Hinderer, Walter et al. (1999). "Serodiagnosis of Epstein-Barr virus infection by using recombinant viral capsid antigen fragments and autologous gene fusion". In: *Journal of Clinical Microbiology* 37.10, pp. 3239–3244. ISSN: 00951137.

Hislop, Andrew D et al. (2007). "Cellular responses to viral infection in humans: lessons from Epstein-Barr virus." In: *Annual review of immunology* 25, pp. 587–617. ISSN: 0732-0582. DOI: 10.1146/annurev.immunol.25.022106.141553.

Hjalgrim, Henrik, Jeppe Friborg, and Mads Melbye (2007). "Chapter 53: The epidemiology of EBV and its association with malignant disease". In: *Human Herpesviruses*. Ed. by Ann Arvin et al. Cambridge: Cambridge University Press. Chap. Chapter 53: The epidemiology of EBV and its association with malignant disease.

Hjalgrim, Henrik et al. (2003). "Characteristics of Hodgkin's lymphoma after infectious mononucleosis." In: *The New England journal of medicine* 349, pp. 1324–1332. ISSN: 0028-4793. DOI: 10.1056/NEJMoa023141.

Hoagland, R. J. (1955). "The Transmission of Infectious Mononucleosis". In: *The American Journal of the Medical Sciences* 3, pp. 229–262.

Honess, R. W. et al. (1989). "Deviations from Expected Frequencies of CpG Dinucleotides in Herpesvirus DNAs May Be Diagnostic of Differences in the States of Their Latent Genomes". In: *Journal of General Virology* 70.4, pp. 837–855.

Horwitz, Charles A. et al. (1981). "Clinical and Laboratory Evaluation of Infants and Children with Epstein-Barr Virus-Induced Infectious Mononucleosis: Report of 32 Patients (Aged 10-48 Months)". In: *Blood* 57.5, pp. 933–938. ISSN: 0006-4971.

Houldcroft, Charlotte J., Mathew A. Beale, and Judith Breuer (2017). "Clinical and biological insights from viral genome sequencing". In: *Nature Reviews Microbiology*. ISSN: 1740-1526. DOI: 10.1038/nrmicro.2016.182.

Hsieh, Pin Pen et al. (2007). "EBV viral load in tumor tissue is an important prognostic indicator for nasal NK/T-cell lymphoma". In: *American Journal of Clinical Pathology* 128.4, pp. 579–584. ISSN: 00029173. DOI: 10.1309/MN4Y8HLQWKD9NB5E.

Hu, L. F. et al. (1991). "Isolation and sequencing of the Epstein-Barr virus BNLF-1 gene (LMP1) from a Chinese nasopharyngeal carcinoma". In: *Journal of General Virology* 72.10, pp. 2399–2409. ISSN: 00221317. DOI: 10.1099/0022-1317-72-10-2399.

Hung, S C, M S Kang, and E Kieff (2001). "Maintenance of Epstein-Barr virus (EBV) oriP-based episomes requires EBV-encoded nuclear antigen-1 chromosome-binding domains, which can be replaced by high-mobility group-I or histone H1." In: *Proceedings of the National Academy of Sciences of the United States of America* 98.4, pp. 1865–1870. ISSN: 0027-8424. DOI: 10.1073/pnas.98.4.1865.

Huson, D H (1998). "SplitsTree: analyzing and visualizing evolutionary data." In: *Bioinformatics (Oxford, England)* 14.1, pp. 68–73. ISSN: 1367-4803. DOI: btb043[pii].

Huson, Daniel H and David Bryant (2006). "Application of phylogenetic networks in evolutionary studies." In: *Molecular biology and evolution* 23.2, pp. 254–67. ISSN: 0737-4038. DOI: 10.1093/molbev/msj030.

Hutt-Fletcher, Lindsey (2016). "The long and complicated relationship between Epstein-Barr virus and epithelial cells". In: *Journal of Virology*, JVI.01677–16. ISSN: 0022-538X. DOI: 10.1128/JVI.01677-16.

Icheva, Vanya et al. (2013). "Adoptive transfer of Epstein-Barr virus (EBV) nuclear antigen 1-specific T cells as treatment for EBV reactivation and lymphoproliferative disorders after allogeneic stem-cell transplantation". In: *Journal of Clinical Oncology* 31.1, pp. 39–48. ISSN: 0732183X. DOI: 10.1200/JCO.2011.39.8495.

Illumina Inc. (2010). *Calling Sequencing SNPs*. Technical Note.

Imai, S et al. (1994). "Gastric carcinoma: monoclonal epithelial malignant cells expressing Epstein-Barr virus latent infection protein." In: *Proceedings of the National Academy of Sciences of the United States of America* 91.19, pp. 9131–5. ISSN: 0027-8424. DOI: 10.1073/pnas.91.19.9131.

Imajoh, Masayuki et al. (2012). "Characterization of Epstein–Barr virus (EBV) BZLF1 gene promoter variants and comparison of cellular gene expression profiles in Japanese patients with infectious mononucleosis, chronic active EBV infection, and EBV-associated hemophagocytic lymphohistiocytosis". In: *Journal of Medical Virology* 84 (6), pp. 940–946. DOI: 10.1002/jmv.23299.

Imig, Jochen et al. (2011). "MicroRNA profiling in Epstein-Barr virus-associated B-cell lymphoma". In: *Nucleic Acids Research* 39.5, pp. 1880–1893. ISSN: 03051048. DOI: 10.1093/nar/gkq1043.

Immune Epitope Database. http://www.iedb.org, accessed last May 2016.

Isobe, Y. (2004). "Epstein-Barr Virus Infection of Human Natural Killer Cell Lines and Peripheral Blood Natural Killer Cells". In: *Cancer Research* 64.6, pp. 2167–2174. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-03-1562.

Iwakiri, Dai and Kenzo Takada (2010). "Role of EBERs in the Pathogenesis of EBV Infection". In: *Advances in Cancer Research* 107, pp. 119 –136. ISSN: 0065-230X. DOI: http://dx.doi.org/10.1016/S0065-230X(10)07004-1.

Iwakiri, Dai et al. (2009). "Epstein-Barr virus (EBV)-encoded small RNA is released from EBV-infected cells and activates signaling from Toll-like receptor 3." In: *The Journal of experimental medicine* 206.10, pp. 2091–9. ISSN: 1540-9538. DOI: 10.1084/jem.20081761.

Izumi, K M and E D Kieff (1997). "The Epstein-Barr virus oncogene product latent membrane protein 1 engages the tumor necrosis factor receptor-associated death domain protein to mediate B lymphocyte growth transformation and activate NF-kappaB." In: *Proceedings of the National Academy of Sciences of the United States of America* 94.23, pp. 12592–12597. ISSN: 0027-8424.

Jabs, Wolfram J et al. (2001). "Normalized Quantification by Real-Time PCR of Epstein-Barr Virus Load in Patients at Risk for Posttransplant Lymphoproliferative Disorders". In: *Journal of Clinical Microbiology*, pp. 564–569. DOI: 10.1128/JCM.39.2.564.

Jakobsson, Mattias and Noah A. Rosenberg (2007). "CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure". In: *Bioinformatics* 23.14, pp. 1801–1806. ISSN: 13674803. DOI: 10.1093/bioinformatics/btm233.

Jia, Wei-Hua et al. (2010). "Traditional Cantonese Diet and Nasopharyngeal Carcinoma Risk: A Large-Scale Case-Control Study in Guangdong, China". In: *BMC cancer* 10, p. 446. ISSN: 1471-2407. DOI: 10.1186/1471-2407-10-446.

Jiang, R, R S Scott, and L M Hutt-Fletcher (2006). "Epstein-Barr virus shed in saliva is high in B-cell-tropic glycoprotein gp42." In: *Journal of virology* 80.14, pp. 7281–3. ISSN: 0022-538X. DOI: 10.1128/JVI.00497-06.

Jin, Yingkang et al. (2010). "Characterization of variants in the promoter of BZLF1 gene of EBV in nonmalignant EBV-associated diseases in Chinese children." In: *Virology journal* 7, p. 92. ISSN: 1743-422X. DOI: 10.1186/1743-422X-7-92.

Johannsen, Eric et al. (2004). "Proteins of purified Epstein-Barr virus." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.46, pp. 16286–91. ISSN: 0027-8424. DOI: 10.1073/pnas.0407320101.

Jombart, Thibaut (2008). "Adegenet: A R package for the multivariate analysis of genetic markers". In: *Bioinformatics* 24.11, pp. 1403–1405. ISSN: 13674803. DOI: 10.1093/bioinformatics/btn129.

Jombart, Thibaut, Sebastien Devillard, and Francois Balloux (2010). "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations". In: *BMC Genetics* 11.1, p. 94. ISSN: 1471-2156. DOI: doi:10.1186/1471-2156-11-94.

Jukes, Thomas H and Charles R Cantor (1969). "Evolution of protein molecules". In: *Mammalian Protein Metabolism*, pp. 21–132. ISBN: 978-1-4832-3211-9. DOI: 10.1016/B978-1-4832-3211-9.50002-4.

Juvonen, E et al. (2003). "High incidence of PTLD after non-T-cell-depleted allogeneic haematopoietic stem cell transplantation as a consequence of intensive immunosuppressive treatment." In: *Bone marrow transplantation* 32.1, pp. 97–102. ISSN: 0268-3369. DOI: 10.1038/sj.bmt.1704089.

Kaiser, C et al. (1999). "The proto-oncogene c-myc is a direct target gene of Epstein-Barr virus nuclear antigen 2." In: *Journal of virology* 73.5, pp. 4481–4484. ISSN: 0022-538X.

Kanda, Teru et al. (2015). "Clustered microRNAs of the Epstein-Barr virus cooperatively downregulate an epithelial cell-specific metastasis suppressor." In: *Journal of virology* 89.5, pp. 2684–97. ISSN: 1098-5514. DOI: 10.1128/JVI.03189-14.

Kang, Dong, Rebecca L. Skalsky, and Bryan R. Cullen (2015). "EBV BART MicroRNAs Target Multiple Pro-apoptotic Cellular Genes to Promote Epithelial Cell Survival." In: *PLoS pathogens* 11.6, e1004979. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1004979.

Kang, Myung-soo and Elliott Kieff (2015). "Epstein – Barr virus latent genes". In: *Experimental & Molecular Medicine* 47.1, e131–16. DOI: 10.1038/emm.2014.84.

Katoh, Kazutaka and Daron M Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." In: *Molecular biology and evolution* 30.4, pp. 772–80. ISSN: 1537-1719. DOI: 10.1093/molbev/mst010.

Kennedy, Gregory, Jun Komano, and Bill Sugden (2003). "Epstein-Barr virus provides a survival factor to Burkitt's lymphomas." In: *Proceedings of the National Academy of Sciences of the United States of America* 100.24, pp. 14269–14274. ISSN: 0027-8424. DOI: 10.1073/pnas.2336099100.

Kenney, Shannon C (2007). "Chapter 25: Reactivation and lytic replication of EBV". In: *Human Herpesviruses*. Ed. by Ann Arvin et al. Cambridge: Cambridge University Press. Chap. Chapter 25: Reactivation and lytic replication of EBV.

Kerkar, Nanda et al. (2010). "The changing face of post-transplant lymphoproliferative disease in the era of molecular EBV monitoring." In: *Pediatric transplantation* 14.4, pp. 504–11. ISSN: 1399-3046. DOI: 10.1111/j.1399-3046.2009.01258.x.

Khan, Gulfaraz et al. (2014). "Global burden of deaths from Epstein-Barr virus attributable malignancies 1990-2010". In: *Infectious Agents and Cancer* 9.1, p. 38. ISSN: 1750-9378. DOI: 10.1186/1750-9378-9-38.

Kienzle, Norbert et al. (1999). "Epstein-Barr Virus-Encoded RK-BARF0 Protein Expression". In: *Journal of Virology* 73.10, pp. 8902–8906. eprint: http://jvi.asm.org/content/73/10/8902.full.pdf+html.

Kim, Sung Min, So Hee Kang, and Won Keun Lee (2006). "Identification of two types of naturally-occurring intertypic recombinants of Epstein-Barr virus". In: *Molecules and Cells* 21.2, pp. 302–307. ISSN: 10168478. DOI: 10.1109/TCOMM.2005.863731.

Kim, Yohan et al. (2012). "Immune epitope database analysis resource". In: *Nucleic Acids Research* 40.W1, pp. 525–530. ISSN: 03051048. DOI: 10.1093/nar/gks438.

Kimura, M (1980). "A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences". In: *Journal of Molecular Evolution* 16.2, pp. 111–120. ISSN: 1098-6596. DOI: 10.1007/bf01731581. arXiv: arXiv:1011.1669v3.

Kimura, Y et al. (2011). "Epidemiological analysis of nasopharyngeal carcinoma in the central region of Japan during the period from 1996 to 2005". In: *Auris Nasus Larynx* 38.2, pp. 244–249. DOI: 10.1016/j.anl.2010.07.006.

Klein, George (1983). "Specific chromosomal translocations and the genesis of B-cell-derived tumors in mice and men". In: *Cell* 32.2, pp. 311 –315. ISSN: 0092-8674. DOI: http://dx.doi.org/10.1016/0092-8674(83)90449-X.

Koboldt, Daniel C. et al. (2012). "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing". In: *Genome Research* 22.3, pp. 568–576. ISSN: 10889051. DOI: 10.1101/gr.129684.111.

Komano, J et al. (1999). "Oncogenic role of Epstein-Barr virus-encoded RNAs in Burkitt's lymphoma cell line Akata." In: *Journal of virology* 73.12, pp. 9827–31. ISSN: 0022-538X.

Kong, Qing Li et al. (2010). "Epstein-barr virus-encoded LMP2A induces an epithelial-mesenchymal transition and increases the number of side population stem-like cancer cells in nasopharyngeal carcinoma". In: *PLoS Pathogens* 6.6. ISSN: 15537366. DOI: 10.1371/journal.ppat.1000940.

Kosakovsky Pond, Sergei L et al. (2006). "GARD: a genetic algorithm for recombination detection." In: *Bioinformatics (Oxford, England)* 22.24, pp. 3096–8. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btl474.

Kroll, Jing et al. (2011). "Lytic and latent EBV gene expression in transplant recipients with and without post-transplant lymphoproliferative disorder". In: *Journal of Clinical Virology* 52.3, pp. 231–235. ISSN: 13866532. DOI: 10.1016/j.jcv.2011.06.013.

Krueger, Felix. *TrimGalore*. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Küppers, R et al. (1994). "Hodgkin disease: Hodgkin and Reed-Sternberg Cells Picked from Histological Sections Show Clonal Immunoglobulin Gene Rearrangements and Appear to be Derived from B Cells at Various Stages of Development". In: *Proceedings of the National Academy of Sciences of the United States of America* 91.November, pp. 10962–10966. ISSN: 0027-8424. DOI: 10.1073/pnas.91.23.10962.

Kurth, Julia et al. (2003). "Epstein-Barr virus-infected B cells expanding in germinal centers of infectious mononucleosis patients do not participate in the germinal center reaction." In: *Proceedings of the National Academy of Sciences of the United States of America* 100.8, pp. 4730–4735. ISSN: 0027-8424. DOI: 10.1073/pnas.2627966100.

Kusano, S and N Raab-Traub (2001). "An Epstein-Barr virus protein interacts with Notch". In: *J Virol* 75.1, pp. 384–395. DOI: 10.1128/JVI.75.1.384-395.2001.

Kuzembayeva, Malika, Mitchell Hayes, and Bill Sugden (2014). "Multiple functions are mediated by the miRNAs of Epstein-Barr virus". In: *Current Opinion in Virology* 7.1, pp. 61–65. ISSN: 18796265. DOI: 10.1016/j.coviro.2014.04.003. arXiv: NIHMS150003.

Kwok, H et al. (2014). "Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsies." In: *Journal of virology*. ISSN: 1098-5514. DOI: 10.1128/JVI.01665-14.

Kwok, Hin et al. (2012). "Genomic sequencing and comparative analysis of Epstein-Barr virus genome isolated from primary nasopharyngeal carcinoma biopsy." In: *PloS one* 7.5, e36939. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0036939.

Kwok, Hin et al. (2015). "Distribution, persistence and interchange of epstein-barr virus strains among PBMC, plasma and saliva of primary infection subjects". In: *PLoS ONE* 10.3. ISSN: 19326203. DOI: 10.1371/journal.pone.0120710.

Lacoste, Vincent et al. (2010). "Genetic diversity and molecular evolution of human and non-human primate Gammaherpesvirinae". In: *Infection, Genetics and Evolution* 10.1, pp. 1 –13. ISSN: 1567-1348. DOI: http://dx.doi.org/10.1016/j.meegid.2009.10.009.

Laichalk, Lauri L and David A Thorley-Lawson (2005). "Terminal differentiation into plasma cells initiates the replicative cycle of Epstein-Barr virus in vivo." In: *Journal of virology* 79.2, pp. 1296–307. ISSN: 0022-538X. DOI: 10.1128/JVI.79.2.1296-1307.2005.

Lassalle, Florent et al. (2016). "Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes". In: *Virus Evolution* 2.1. DOI: 10.1093/ve/vew017. eprint: http://ve.oxfordjournals.org/content/2/1/vew017.full.pdf.

Lee, Hye Seung et al. (2004). "Epstein-Barr Virus-Positive Gastric Carcinoma Has a Distinct Protein Expression Profile in Comparison with Epstein-Barr Virus-Negative Carcinoma". In: *Clinical Cancer Research* 10.5, pp. 1698–1705. ISSN: 10780432. DOI: 10.1158/1078-0432.CCR-1122-3.

Lee, Ju-Han et al. (2009). "Clinicopathological and molecular characteristics of Epstein-Barr virus-associated gastric carcinoma: a meta-analysis." In: *Journal of gastroenterology and hepatology* 24.3, pp. 354–65. ISSN: 1440-1746. DOI: 10.1111/j.1440-1746.2009.05775.x.

Lei, Haiyan et al. (2013). "Identification and characterization of EBV genomes in spontaneously immortalized human peripheral blood B lymphocytes by NGS technology." In: *BMC genomics* 14, p. 804. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-804.

Lei, Haiyan et al. (2015). "Epstein-Barr virus from Burkitt Lymphoma biopsies from Africa and South America share novel LMP-1 promoter and gene variations." In: *Scientific reports* 5.October, p. 16706. ISSN: 2045-2322. DOI: 10.1038/srep16706.

Lemey, P., Salemi, M., and Vandamme, A. M. (2009). *The Phylogentic Handbook. A Practical Approach to phylogentic analysis and hypothesis testing*. P. 723. ISBN: 978-0-521-73071-6.

Li, Ang et al. (2005). "Transcriptional expression of RPMS1 in nasopharyngeal carcinoma and its oncogenic potential". In: *Cell Cycle* 4.2, pp. 304–309. ISSN: 15384101. DOI: 10.4161/cc.4.2.1416.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." In: *Bioinformatics (Oxford, England)* 25.14, pp. 1754–60. ISSN: 1367-4811. DOI: 10.1093/\\bioinformatics/btp324.

Li, Heng et al. (2009). "The Sequence Alignment/Map format and SAMtools." In: *Bioinformatics (Oxford, England)* 25.16, pp. 2078–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352.

Li, S N, Y S Chang, and S T Liu (1996). "Effect of a 10-amino acid deletion on the oncogenic activity of latent membrane protein 1 of Epstein-Barr virus." In: *Oncogene* 12.10, pp. 2129–2135. ISSN: 0950-9232 (Print).

Lin, Xiaochen et al. (2015). "The Epstein-Barr Virus BART miRNA Cluster of the M81 Strain Modulates Multiple Functions in Primary B Cells". In: *PLoS Pathogens* 11.12. ISSN: 15537374. DOI: 10.1371/journal.ppat.1005344.

Lin, Zhen et al. (2012). "Whole genome sequencing of the Akata and Mutu Epstein-Barr virus (EBV) strains." In: *Journal of virology*. ISSN: 1098-5514. DOI: 10.1128/JVI.02517-12.

Ling, P D et al. (2003). "The dynamics of herpesvirus and polyomavirus reactivation and shedding in healthy adults: a 14-month longitudinal study". In: *J Infect Dis* 187.10, pp. 1571–1580. ISSN: 0022-1899. DOI: 10.1086/374739.

Liu, Jia et al. (2014). "Extensive recombination due to heteroduplexes generates large amounts of artificial gene fragments during pcr". In: *PLoS ONE* 9.9. ISSN: 19326203. DOI: 10.1371/journal.pone.0106658.

Liu, Pan et al. (2011). "Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology." In: *Journal of virology* 85.21, pp. 11291–9. ISSN: 1098-5514. DOI: 10.1128/JVI.00823-11.

Liu, Ying et al. (2016). "Genome-wide analysis of Epstein-Barr virus (EBV) isolated from EBV-associated gastric carcinoma (EBVaGC)." In: *Oncotarget* 7.4, pp. 4903–14. ISSN: 1949-2553. DOI: 10.18632/oncotarget.6751.

Lo, Angela Kwok Fung et al. (2007). "Modulation of LMP1 protein expression by EBV-encoded microRNAs." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.41, pp. 16164–9. ISSN: 0027-8424. DOI: 10.1073/pnas.0702896104.

Long, Heather M et al. (2013). "MHC II tetramers visualize human CD4+ T cell responses to Epstein-Barr virus infection and demonstrate atypical kinetics of the nuclear antigen EBNA1 response." In: *The Journal of experimental medicine* 210.5, pp. 933–49. ISSN: 1540-9538. DOI: 10.1084/jem.20121437.

Lorenzetti, M. A. et al. (2014). "Epstein-Barr virus BZLF1 gene polymorphisms: Malignancy related or geographically distributed variants?" In: *Clinical Microbiology and Infection* 20.11, O861–O869. ISSN: 14690691. DOI: 10.1111/1469-0691.12631.

Lorenzetti, Mario Alejandro et al. (2009). "Epstein-Barr virus BZLF1 gene promoter variants in pediatric patients with acute infectious mononucleosis: Its comparison with pediatric lymphomas". In: *Journal of Medical Virology* 81.11, pp. 1912–1917. ISSN: 01466615. DOI: 10.1002/jmv.21616.

Lorenzetti, Mario Alejandro et al. (2012). "Distinctive Epstein-Barr virus variants associated with benign and malignant pediatric pathologies: LMP1 sequence characterization and linkage with other viral gene polymorphisms". In: *Journal of Clinical Microbiology* 50.3, pp. 609–618. ISSN: 00951137. DOI: 10.1128/JCM.05778-11.

Lu, Jean et al. (2006). "Syk tyrosine kinase mediates Epstein-Barr virus latent membrane protein 2A-induced cell migration in epithelial cells." In: *The Journal of biological chemistry* 281.13, pp. 8806–8814. ISSN: 0021-9258. DOI: 10.1074/jbc.M507305200.

Lundegaard, Claus et al. (2008). "NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11." In: *Nucleic acids research* 36.Web Server issue, pp. 509–512. ISSN: 13624962. DOI: 10.1093/nar/gkn202.

Luo, Bing et al. (2011). "Sequence variation of Epstein-Barr virus (EBV) BZLF1 gene in EBV-associated gastric carcinomas and nasopharyngeal carcinomas in Northern China". In: *Microbes and Infection* 13.8-9, pp. 776–782. ISSN: 12864579. DOI: 10.1016/j.micinf.2011.04.002.

Luo, Wen-Juan et al. (2004). "Epstein-Barr virus is integrated between REL and BCL-11A in American Burkitt lymphoma cell line (NAB-2)." In: *Laboratory investigation; a journal of technical methods and pathology* 84.9, pp. 1193–9. ISSN: 0023-6837. DOI: 10.1038/labinvest.3700152.

Ma, Shi-Dong et al. (2011). "A new model of Epstein-Barr virus infection reveals an important role for early lytic viral protein expression in the development of lymphomas." In: *Journal of virology* 85.1, pp. 165–77. ISSN: 1098-5514. DOI: 10.1128/JVI.01512-10.

Mai, Shi Juan et al. (2007). "Functional advantage of NPC-related V-val subtype of Epstein-Barr virus nuclear antigen 1 compared with prototype in epithelial cell line". In: *Oncology Reports* 17.1, pp. 141–146. ISSN: 1021335X.

Mai, Shi-Juan et al. (2010). "The enhanced transcriptional activity of the V-val subtype of Epstein-Barr virus nuclear antigen 1 in epithelial cell lines." In: *Oncology reports* 23.5, pp. 1417–1424. ISSN: 1791-2431 (Electronic).

Mainou, Bernardo a and Nancy Raab-Traub (2006). "LMP1 strain variants: biological and molecular properties." In: *Journal of virology* 80.13, pp. 6458–6468. ISSN: 0022-538X. DOI: 10.1128/JVI.00135-06.

Mancao, C et al. (2005). "Rescue of "crippled" germinal center B cells from apoptosis by Epstein-Barr virus". In: *Blood* 106.13, pp. 4339–4344. DOI: 2005-06-2341.

Marquitz, Aron R et al. (2013). "Expression profile of microRNAs in EBV infected AGS gastric carcinoma cells." In: *Journal of virology* 88.2, pp. 1389–1393. ISSN: 1098-5514. DOI: 10.1128/JVI.02662-13.

Martini, Maurizio et al. (2007). "Characterization of variants in the promoter of EBV gene BZLF1 in normal donors, HIV-positive patients and in AIDS-related lymphomas". In: *Journal of Infection* 54.3, pp. 298–306. ISSN: 01634453. DOI: 10.1016/j.jinf.2006.04.015.

Matsuo, T et al. (1984). "Persistence of the entire Epstein-Barr virus genome integrated into human lymphocyte DNA." In: *Science (New York, N.Y.)* 226.4680, pp. 1322–1325. ISSN: 0036-8075 (Print). DOI: 10.1126/science.6095452.

Mautner, Josef and Georg W Bornkamm (2012). "The role of virus-specific CD4+ T cells in the control of Epstein-Barr virus infection." In: *European journal of cell biology* 91.1, pp. 31–5. ISSN: 1618-1298. DOI: 10.1016/j.ejcb.2011.01.007.

McGeoch, Duncan J., Derek Gatherer, and Aidan Dolan (2005). "On phylogenetic relationships among major lineages of the Gammaherpesvirinae". In: *Journal of General Virology* 86.2, pp. 307–316.

McGeoch, Duncan J. et al. (1995). "Molecular Phylogeny and Evolutionary Timescale for the Family of Mammalian Herpesviruses". In: *Journal of Molecular Biology* 247.3, pp. 443 –458. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1006/jmbi.1995.0152.

Meckes, D. G. et al. (2010). "Human tumor virus utilizes exosomes for intercellular communication". In: *Proceedings of the National Academy of Sciences* 107.47, pp. 20370–20375. ISSN: 0027-8424. DOI: 10.1073/pnas.1014194107.

Melnikov, Alexandre et al. (2011). "Hybrid selection for sequencing pathogen genomes from clinical samples". In: *Genome Biology* 12.8, R73. ISSN: 1465-6906. DOI: 10.1186/gb-2011-12-8-r73.

Midgley, R. S. et al. (2000). "Novel Intertypic Recombinants of Epstein-Barr Virus in the Chinese Population". In: *Journal of Virology* 74.3, pp. 1544–1548. ISSN: 0022-538X. DOI: 10.1128/JVI.74.3.1544-1548.2000.

Miller, W E et al. (1994). "Sequence variation in the Epstein-Barr virus latent membrane protein 1." In: *The Journal of general virology* 75 (Pt 10, pp. 2729–40. ISSN: 0022-1317.

Milner, A E et al. (1993). "Apoptosis in Burkitt lymphoma cells is driven by c-myc." In: *Oncogene* 8.12, pp. 3385–3391. ISSN: 0950-9232 (Print).

Mosser, David M. and Xia Zhang (2008). *Interleukin-10: New perspectives on an old cytokine*. DOI: 10.1111/j.1600-065X.2008.00706.x. arXiv: NIHMS150003.

Nakai, Hidetaka et al. (2012). "Host factors associated with the kinetics of Epstein-Barr virus DNA load in patients with primary Epstein-Barr virus infection". In: *Microbiology and Immunology* 56.2, pp. 93–98. ISSN: 03855600. DOI: 10.1111/j.1348-0421.2011.00410.x.

Nanbo, Asuka, Hironori Yoshiyama, and Kenzo Takada (2005). "Epstein-Barr Virus-Encoded Poly(A)- RNA Confers Resistance to Apoptosis Mediated through Fas by Blocking the PKR Pathway in Human Epithelial Intestine 407 Cells". In: *Journal of Virology* 79.19,

pp. 12280–12285. DOI: 10.1128/JVI.79.19.12280-12285.2005. eprint: http://jvi.asm.org/content/79/19/12280.full.pdf+html.

Nanbo, Asuka et al. (2002). "Epstein-Barr virus RNA confers resistance to interferon-alpha-induced apoptosis in Burkitt's lymphoma." In: *The EMBO journal* 21.5, pp. 954–65. ISSN: 0261-4189. DOI: 10.1093/emboj/21.5.954.

Nei, M (1987). *Molecular Evolutionary Genetics*. Vol. 17, p. 512. ISBN: 0231063202.

Nei, Masatoshi and Wen-Hsiung Li (1979). "Mathematical model for studying genetic variation in terms of restriction endonucleases." In: *Proceedings of the National Academy of Sciences of the United States of America* 76.10, pp. 5269–5273. ISSN: 0027-8424. DOI: 10.1073/pnas.76.10.5269.

Nemerow, G R et al. (1987). "Identification of gp350 as the viral glycoprotein mediating attachment of Epstein-Barr virus (EBV) to the EBV/C3d receptor of B cells: sequence homology of gp350 and C3 complement fragment C3d." In: *Journal of Virology* 61.5, pp. 1416–1420. eprint: http://jvi.asm.org/content/61/5/1416.full.pdf+html.

Niedobitek, Gerald et al. (1997). "Epstein-Barr virus (EBV) infection in infectious mononucleosis: Virus latency, replication and phenotype of EBV-infected cells". In: *Journal of Pathology* 182.2, pp. 151–159. ISSN: 00223417. DOI: 10.1002/(SICI)1096-9896(199706)182:2<151::AID-PATH824>3.0.CO;2-3.

Nielsen, Morten et al. (2003). "Reliable prediction of T-cell epitopes using neural networks with novel sequence representations." In: *Protein science : a publication of the Protein Society* 12.5, pp. 1007–1017. ISSN: 0961-8368. DOI: 10.1110/ps.0239403.

Nijland, Marieke L et al. (2016). "Epstein-Barr Virus-Positive Posttransplant Lymphoproliferative Disease After Solid Organ Transplantation: Pathogenesis, Clinical Manifestations, Diagnosis, and Management." In: *Transplantation direct* 2.1, e48. ISSN: 2373-8731. DOI: 10.1097/TXD.0000000000000557.

Odumade, Oludare a, Kristin a Hogquist, and Henry H Balfour (2011). "Progress and problems in understanding and managing primary Epstein-Barr virus infections." In: *Clinical microbiology reviews* 24.1, pp. 193–209. ISSN: 1098-6618. DOI: 10.1128/CMR.00044-10.

Oertel, Stephan H K et al. (2005). "Effect of anti-CD 20 antibody rituximab in patients with post-transplant lymphoproliferative disorder (PTLD)". In: *American Journal of Transplantation* 5.12, pp. 2901–2906. ISSN: 16006135. DOI: 10.1111/j.1600-6143.2005.01098.x.

Ordonez, Paula et al. (2011). "Identification of the distinctive type i/XhoI+ strain of epstein-barr virus in gastric carcinoma in Peru". In: *Anticancer Research* 31.10, pp. 3607–3613. ISSN: 02507005.

Pakpoor, Julia et al. (2013). "The risk of developing multiple sclerosis in individuals seronegative for Epstein-Barr virus: a meta-analysis". In: *Multiple Sclerosis Journal* 19.2, pp. 162–166. ISSN: 1352-4585. DOI: 10.1177/1352458512449682.

Palser, Anne L. et al. (2015). "Genome Diversity of Epstein-Barr Virus from Multiple Tumour Types and Normal Infection". In: *Journal of Virology* March, JVI.03614–14. ISSN: 0022-538X. DOI: 10.1128/JVI.03614-14.

Pan, Shih Hsuan et al. (2009). "Epstein-Barr virus nuclear antigen 2 disrupts mitotic checkpoint and causes chromosomal instability". In: *Carcinogenesis* 30.2, pp. 366–375. ISSN: 01433334. DOI: 10.1093/carcin/bgn291.

Panikkar, Archana et al. (2015). "Impaired Epstein-Barr Virus-Specific Neutralizing Antibody Response during Acute Infectious Mononucleosis Is Coincident with Global B-Cell Dysfunction." In: *Journal of virology* 89.17, pp. 9137–9141. ISSN: 1098-5514 (Electronic). DOI: 10.1128/JVI.01293-15.

Paradis, E., J. Claude, and K. Strimmer (2004). "APE: Analyses of Phylogenetics and Evolution in R language". In: *Bioinformatics* 20.2, pp. 289–290. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg412.

Paradis, Emmanuel (2010). "pegas: an R package for population genetics with an integrated-modular approach." In: *Bioinformatics (Oxford, England)* 26.3, pp. 419–20. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp696.

Pavlopoulos, Georgios A et al. (2011). "Using graph theory to analyze biological networks." In: *BioData mining* 4.1, p. 10. ISSN: 1756-0381. DOI: 10.1186/1756-0381-4-10.

Pender, Michael P. (2003). *Infection of autoreactive B lymphocytes with EBV, causing chronic autoimmune diseases*. DOI: 10.1016/j.it.2003.09.005.

Pender, Michael P and Scott R Burrows (2014). "Epstein-Barr virus and multiple sclerosis: potential opportunities for immunotherapy." In: *Clinical & translational immunology* 3.10, e27. ISSN: 2050-0068. DOI: 10.1038/cti.2014.25.

Pérez-Losada, Marcos et al. (2015). *Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences*. DOI: 10.1016/j.meegid.2014.12.022.

Peters, Bjoern and Alessandro Sette (2005). "Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method." In: *BMC bioinformatics* 6, p. 132. ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-132.

Pfeffer, Sébastien et al. (2004). "Identification of virus-encoded microRNAs." In: *Science (New York, N.Y.)* 304.5671, pp. 734–6. ISSN: 1095-9203. DOI: 10.1126/science.1096781. arXiv: 0208024 [gr-qc].

Picard. http://broadinstitute.github.io/picard.

Pinard, Robert et al. (2006). "Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing." In: *BMC genomics* 7, p. 216. ISSN: 1471-2164. DOI: 10.1186/1471-2164-7-216.

Piovan, Erich et al. (2005). "Chemokine receptor expression in EBV-associated lymphoproliferation in hu/SCID mice: Implications for CXCL12/CXCR4 axis in lymphoma generation". In: *Blood* 105.3, pp. 931–939. ISSN: 00064971. DOI: 10.1182/blood-2004-03-0799.

Pitman, S.D. et al. (2006). "Hodgkin lymphoma-like posttransplant lymphoproliferative disorder (HL-like PTLD) simulates monomorphic B-cell PTLD both clinically and

pathologically". In: *American Journal of Surgical Pathology* 30.4, pp. 470–476. ISSN: 01475185. DOI: 10.1097/00000478-200604000-00007.

Pond, Sergei L Kosakovsky, Simon D W Frost, and Spencer V Muse (2005). "HyPhy: hypothesis testing using phylogenies." In: *Bioinformatics (Oxford, England)* 21.5, pp. 676–9. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti079.

Prang, N S et al. (1997). "Lytic replication of Epstein-Barr virus in the peripheral blood: analysis of viral gene expression in B lymphocytes during infectious mononucleosis and in the normal carrier state." In: *Blood* 89.5, pp. 1665–1677. ISSN: 0006-4971.

Pritchard, J K, M Stephens, and P Donnelly (2000). "Inference of population structure using multilocus genotype data." In: *Genetics* 155.2, pp. 945–59. ISSN: 0016-6731. DOI: 10.1111/j.1471-8286.2007.01758.x.

Qiu, Jin et al. (2015). "The Epstein-Barr Virus Encoded BART miRNAs Potentiate Tumor Growth In Vivo". In: *PLOS Pathogens* 11.1, pp. 1–22. DOI: 10.1371/journal.ppat.1004561.

Qu, Lirong et al. (2000). "Epstein-Barr Virus Gene Expression in the Peripheral Blood of Transplant Recipients with Persistent Circulating Virus Loads". In: *The Journal of Infectious Diseases* 182.4, pp. 1013–1021. ISSN: 0022-1899. DOI: 10.1086/315828.

Quan, Timothy E et al. (2010). "Epstein-Barr virus promotes interferon-alpha production by plasmacytoid dendritic cells." In: *Arthritis and rheumatism* 62.6, pp. 1693–1701. ISSN: 1529-0131 (Electronic). DOI: 10.1002/art.27408.

QUASR. http://sourceforge.net/projects/quasr/.

Quinn, Laura L et al. (2016). "The Missing Link in Epstein-Barr Virus Immune Evasion: the BDLF3 Gene Induces Ubiquitination and Downregulation of Major Histocompatibility Complex Class I (MHC-I) and MHC-II." In: *Journal of virology* 90.1, pp. 356–67. ISSN: 1098-5514. DOI: 10.1128/JVI.02183-15.

Raab-Traub, Nancy and Kathy Flynn (1986). "The structure of the termini of the Epstein-Barr virus as a marker of clonal cellular proliferation". In: *Cell* 47.6, pp. 883–889. ISSN: 00928674. DOI: 10.1016/0092-8674(86)90803-2.

Ran, F Ann et al. (2013). "Genome engineering using the CRISPR-Cas9 system". In: *Nat. Protocols* 8.11, pp. 2281–2308. ISSN: 1754-2189. DOI: 10.1038/nprot.2013.143\rhttp://www.nature.com/nprot/journal/v8/n11/abs/nprot.2013.143.html#supplementary-information. arXiv: NIHMS150003.

Renzette, N et al. (2014). "Epstein-Barr virus latent membrane protein 1 genetic variability in peripheral blood B cells and oropharyngeal fluids". In: *J Virol* 88.7, pp. 3744–3755. DOI: 10.1128/JVI.03378-13.

Rickinson, a B, L S Young, and M Rowe (1987). "Influence of the Epstein-Barr virus nuclear antigen EBNA 2 on the growth phenotype of virus-transformed B cells." In: *Journal of virology* 61.5, pp. 1310–1317. ISSN: 0022-538X.

Rivailler, Pierre, Young-gyu Cho, and Fred Wang (2002). "Complete Genomic Sequence of an Epstein-Barr Virus-Related Herpesvirus Naturally Infecting a New World Primate: a Defining Point in the Evolution of Oncogenic Lymphocryptoviruses". In: *Journal of*

Virology 76.23, pp. 12055–12068. DOI: 10.1128/JVI.76.23.12055-12068.2002. eprint: http://jvi.asm.org/content/76/23/12055.full.pdf+html.

Rosa, M D et al. (1981). "Striking similarities are exhibited by two small Epstein-Barr virus-encoded ribonucleic acids and the adenovirus-associated ribonucleic acids VAI and VAII." In: *Molecular and cellular biology* 1.9, pp. 785–96. ISSN: 0270-7306. DOI: 10.1128/MCB.1.9.785.

Rose, C. et al. (2001). "Pediatric solid-organ transplant recipients carry chronic loads of Epstein-Barr virus exclusively in the immunoglobulin D-negative B-cell compartment". In: *Journal of Clinical Microbiology* 39.4, pp. 1407–1415. ISSN: 00951137. DOI: 10.1128/JCM.39.4.1407-1415.2001.

Rose, Camille et al. (2002). "Detection of Epstein-Barr virus genomes in peripheral blood B cells from solid-organ transplant recipients by fluorescence in situ hybridization". In: *Journal of Clinical Microbiology* 40.7, pp. 2533–2544. ISSN: 00951137. DOI: 10.1128/JCM.40.7.2533-2544.2002.

Rosenberg, Noah A. (2004). "DISTRUCT: A program for the graphical display of population structure". In: *Molecular Ecology Notes* 4.1, pp. 137–138. ISSN: 14718278. DOI: 10.1046/j.1471-8286.2003.00566.x.

Roskrow, M A et al. (1998). "Epstein-Barr virus (EBV)-specific cytotoxic T lymphocytes for the treatment of patients with EBV-positive relapsed Hodgkin's disease". In: *Blood* 91.8, pp. 2925–2934. ISSN: 0006-4971; 0006-4971.

Rowe, M et al. (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." In: *Journal of virology* 63.3, pp. 1031–9. ISSN: 0022-538X.

Ruiss, Romana et al. (2011). "A virus-like particle-based Epstein-Barr virus vaccine." In: *Journal of virology* 85.24, pp. 13105–13. ISSN: 1098-5514. DOI: 10.1128/JVI.05598-11.

Rymo, L (1979). "Identification of transcribed regions of Epstein-Barr virus DNA in Burkitt lymphoma-derived cells." In: *Journal of virology* 32.1, pp. 8–18. ISSN: 0022538X.

Sacaze, Céline et al. (2001). "Tissue specific distribution of Epstein-Barr virus (EBV) BZLF1 gene variants in nasopharyngeal carcinoma (NPC) bearing patients". In: *Virus Research* 81.1-2, pp. 133–142. ISSN: 01681702. DOI: 10.1016/S0168-1702(01)00376-8.

Saha, Abhik et al. (2012). "E2F1 mediated apoptosis induced by the DNA damage response is blocked by EBV nuclear antigen 3C in lymphoblastoid cells". In: *PLoS Pathogens* 8.3. ISSN: 15537366. DOI: 10.1371/journal.ppat.1002573.

Saito, Shinichi et al. (2013). "Epstein-Barr virus deubiquitinase downregulates TRAF6-mediated NF-κB signaling during productive replication." In: *Journal of virology* 87.7, pp. 4060–70. ISSN: 1098-5514. DOI: 10.1128/JVI.02020-12.

Sample, J et al. (1990). "Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA-3B, and EBNA-3C genes." In: *Journal of virology* 64.9, pp. 4084–92. ISSN: 0022-538X.

Sandvej, Kristian, Xiao G. Zhou, and Stephen Hamilton-Dutoit (2000). "EBNA-1 sequence variation in Danish and Chinese EBV-associated tumours: Evidence for geographical

polymorphism but not for tumour-specific subtype restriction". In: *Journal of Pathology* 191.2, pp. 127–131. ISSN: 00223417. DOI: 10.1002/(SICI)1096-9896(200006)191:2<127::AID-PATH614>3.0.CO;2-E.

Santón, Almudena et al. (2011). "High frequency of co-infection by Epstein-Barr virus types 1 and 2 in patients with multiple sclerosis." In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 17.11, pp. 1295–300. ISSN: 1477-0970. DOI: 10.1177/1352458511411063.

Santpere, Gabriel et al. (2014). "Genome-wide analysis of wild-type epstein-barr virus genomes derived from healthy individuals of the 1000 genomes project". In: *Genome Biology and Evolution* 6.4, pp. 846–860. ISSN: 17596653. DOI: 10.1093/gbe/evu054.

Schneider, B G et al. (2000). "Loss of p16/CDKN2A tumor suppressor protein in gastric adenocarcinoma is associated with Epstein-Barr virus and anatomic location in the body of the stomach." In: *Human pathology* 31.1, pp. 45–50. ISSN: 0046-8177.

Scholle, Frank, Katharine M Bendt, and Nancy Raab-Traub (2000). "Epstein-Barr Virus LMP2A Transforms Epithelial Cells, Inhibits Cell Differentiation, and Activates Akt". In: *JOURNAL OF VIROLOGY* 74.22, pp. 10681–10689. ISSN: 0022-538X. DOI: 10.1128/JVI.74.22.10681-10689.2000.Updated.

Sculley, T. B. et al. (1990). "Coinfection with A and B-Type Epstein-Barr Virus in Human Immunodeficiency Virus-Positive Subjects". In: *Journal of Infectious Diseases* 162.3, pp. 643–648. ISSN: 0022-1899. DOI: 10.1093/infdis/162.3.642.

Serafini, Barbara et al. (2007). "Dysregulated Epstein-Barr virus infection in the multiple sclerosis brain." In: *The Journal of experimental medicine* 204.12, pp. 2899–912. ISSN: 1540-9538. DOI: 10.1084/jem.20071030.

Sette, A et al. (1994). "The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes." In: *Journal of immunology (Baltimore, Md. : 1950)* 153.12, pp. 5586–92. ISSN: 0022-1767. DOI: 153:5586-92.

Shen, Zhi Chao et al. (2015). "High prevalence of the EBER variant EB-8m in endemic nasopharyngeal carcinomas". In: *PLoS ONE* 10.3. ISSN: 19326203. DOI: 10.1371/journal.pone.0121420.

Sidney, John et al. (2008). "Quantitative peptide binding motifs for 19 human and mouse MHC class i molecules derived using positional scanning combinatorial peptide libraries". In: *Immunome Research* 4.1, pp. 1–14. ISSN: 17457580. DOI: 10.1186/1745-7580-4-2.

Sitki-Green, Diane, Mary Covington, and Nancy Raab-Traub (2003). "Compartmentalization and Transmission of Multiple Epstein-Barr Virus Strains in Asymptomatic Carriers Compartmentalization and Transmission of Multiple Epstein-Barr Virus Strains in Asymptomatic Carriers". In: *Journal of Virology* 77.3, pp. 1840–1847. ISSN: 0022-538X. DOI: 10.1128/JVI.77.3.1840.

Sitki-Green, Diane L et al. (2004). "Biology of Epstein-Barr virus during infectious mononucleosis." In: *The Journal of infectious diseases* 189.3, pp. 483–92. ISSN: 0022-1899. DOI: 10.1086/380800.

Skalska, Lenka et al. (2013). "Induction of p16INK4a Is the Major Barrier to Proliferation when Epstein-Barr Virus (EBV) Transforms Primary B Cells into Lymphoblastoid Cell

Lines". In: *PLoS Pathogens* 9.2. ISSN: 15537366. DOI: 10.1371/journal.ppat.1003187.

Smith, Corey and Rajiv Khanna (2012). "A new approach for cellular immunotherapy of nasopharyngeal carcinoma." In: *Oncoimmunology* 1.8, pp. 1440–1442. ISSN: 2162-4011. DOI: 10.4161/onci.21286.

Sokal, Etienne M. et al. (2007). "Recombinant gp350 Vaccine for Infectious Mononucleosis: A Phase 2, Randomized, Double- Blind, Placebo-Controlled Trial to Evaluate the Safety, Immunogenicity, and Efficacy of an Epstein- Barr Virus Vaccine in Healthy Young Adults". In: *The Journal of Infectious Diseases* 196.12, p. 1749. DOI: 10.1086/523813. eprint: /oup/backfile/content_public/journal/jid/196/12/10.1086/523813/2/196-12-1749.pdf.

Spriggs, M K et al. (1996). "The extracellular domain of the Epstein-Barr virus BZLF2 protein binds the HLA-DR beta chain and inhibits antigen presentation." In: *Journal of Virology* 70.8, pp. 5557–63. eprint: http://jvi.asm.org/content/70/8/5557.full.pdf+html.

Sprunt, T P and F A Evans (1920). "Mononuclear leucocytosis in reaction to acute infections ('infectious mononucleosis')." In: *Johns Hopkins Hosp Bull* 31, pp. 410–417.

Srivastava, G et al. (2000). "Coinfection of multiple strains of Epstein-Barr virus in immunocompetent normal individuals: reassessment of the viral carrier state." In: *Blood* 95.7, pp. 2443–2445. ISSN: 00064971.

Straathof, Karin C M et al. (2005). "Treatment of nasopharyngeal carcinoma with Epstein-Barr virus-specific T lymphocytes". In: *Blood* 105.5, pp. 1898–1904. ISSN: 00064971. DOI: 10.1182/blood-2004-07-2975.

Swaminathan, S, B Tomkinson, and E Kieff (1991). "Recombinant Epstein-Barr virus with small RNA (EBER) genes deleted transforms lymphocytes and replicates in vitro." In: *Proceedings of the National Academy of Sciences of the United States of America* 88.4, pp. 1546–50. ISSN: 0027-8424. DOI: 10.1073/pnas.88.4.1546.

Swerdlow, S.H. et al. (2008). *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. Lyon, France*. Vol. 4th, p. 326. ISBN: 9789283224310. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Swerdlow, Steven H. (2007). "T-cell and NK-cell posttransplantation lymphoproliferative disorders". In: *American Journal of Clinical Pathology*. Vol. 127. 6, pp. 887–895. DOI: 10.1309/LYXN3RGF7D7KPYG0.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism". In: *Genetics* 123.3, pp. 585–595. ISSN: 00166731. DOI: PMC1203831.

Takahashi, Michiaki et al. (1974). "Live Vaccine used to prevent the spread of varicella in children in hospital". In: *The Lancet* 304.7892, pp. 1288–1290. ISSN: 01406736. DOI: 10.1016/S0140-6736(74)90144-5.

Tamura, K and M Nei (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." In: *Molecular biology and evolution* 10.3, pp. 512–26. ISSN: 0737-4038. DOI: 10.1093/molbev/msl149.

Tamura, Koichiro et al. (2013). "MEGA6: Molecular Evolutionary Genetics Analysis version 6.0." In: *Molecular biology and evolution* 30.12, pp. 2725–9. ISSN: 1537-1719. DOI: 10.1093/molbev/mst197.

Tanaka, Michiko et al. (1999). "Sequence variations of Epstein-Barr virus LMP2A gene in gastric carcinoma in Japan". In: *Virus Genes* 19.2, pp. 103–111. ISSN: 09208569. DOI: 10.1023/A:1008171006400.

Tao, Q et al. (1998). "The Epstein-Barr virus major latent promoter Qp is constitutively active, hypomethylated, and methylation sensitive." In: *Journal of virology* 72.9, pp. 7075–7083. ISSN: 0022-538X.

Tavaré, S (1986). *Some probabilistic and statistical problems in the analysis of DNA sequences*.

Tellam, Judy T. et al. (2012). "Messenger RNA Sequence Rather than Protein Sequence Determines the Level of Self-synthesis and Antigen Presentation of the EBV-encoded Antigen, EBNA1". In: *PLoS Pathogens* 8.12. ISSN: 15537366. DOI: 10.1371/journal.ppat.1003112.

Thacker, Evan L., Fariba Mirzaei, and Alberto Ascherio (2006). *Infectious mononucleosis and risk for multiple sclerosis: A meta-analysis*. DOI: 10.1002/ana.20820.

Thoendel, Matthew et al. (2017). "Impact of Contaminating DNA in Whole Genome Amplification Kits Used for Metagenomic Shotgun Sequencing for Infection Diagnosis". In: *Journal of Clinical Microbiology*, JCM.02402–16. ISSN: 0095-1137. DOI: 10.1128/JCM.02402-16.

Thorley-Lawson, D A (2001). "Epstein-Barr virus: exploiting the immune system." In: *Nature reviews. Immunology* 1.1, pp. 75–82. ISSN: 1474-1733. DOI: 10.1038/35095584.

Thorley-Lawson, David A. and Andrew Gross (2004). "Persistence of the Epstein–Barr Virus and the Origins of Associated Lymphomas". In: *The New England journal of medicine* 350.13, pp. 1328–1337. ISSN: 0028-4793. DOI: 10.1056/NEJMra032015.

Thornburg, N J, S Kusano, and N Raab-Traub (2004). "Identification of Epstein-Barr virus RK-BARF0-interacting proteins and characterization of expression pattern." In: *J Virol* 78.23, pp. 12848–12856. ISSN: 0022-538X. DOI: 10.1128/JVI.78.23.12848-12856.2004.

Tierney, Rosemary J et al. (2006). "Multiple Epstein-Barr virus strains in patients with infectious mononucleosis: comparison of ex vivo samples with in vitro isolates by use of heteroduplex tracking assays." In: *The Journal of infectious diseases* 193.2, pp. 287–97. ISSN: 0022-1899. DOI: 10.1086/498913.

Tierney, Rosemary J. et al. (2011). "Epstein-Barr Virus BamHI W Repeat Number Limits EBNA2/EBNA-LP Coexpression in Newly Infected B Cells and the Efficiency of B-Cell Transformation: a Rationale for the Multiple W Repeats in Wild-Type Virus Strains". In: *Journal of Virology* 85.23, pp. 12362–12375. DOI: 10.1128/JVI.06059-11. eprint: http://jvi.asm.org/content/85/23/12362.full.pdf+html.

Tobollik, Stephanie et al. (2006). "Epstein-Barr virus nuclear antigen 2 inhibits AID expression during EBV-driven B-cell growth". In: *Blood* 108.12, pp. 3859–3864. ISSN: 00064971. DOI: 10.1182/blood-2006-05-021303.

Tomkinson, B, E Robertson, and E Kieff (1993). "Epstein-Barr virus nuclear proteins EBNA-3A and EBNA-3C are essential for B-lymphocyte growth transformation." In: *Journal of virology* 67.4, pp. 2014–2025. ISSN: 0022-538X.

Tong, Joanna H. M. et al. (2003). "Re: Discrete Alterations in the BZLF1 Promoter in Tumor and Non-Tumor-Associated Epstein–Barr Virus". In: *Journal of the National Cancer Institute* 95.13, pp. 1008–1009. DOI: 10.1093/jnci/95.13.1008. eprint: http://jnci.oxfordjournals.org/content/95/13/1008.full.pdf+html.

Tsai, Ming-Han et al. (2013). "Spontaneous lytic replication and epitheliotropism define an epstein-barr virus strain found in carcinomas." In: *Cell reports* 5.2, pp. 458–70. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2013.09.012.

Tschochner, Monika et al. (2016). "Identifying patient-specific Epstein-Barr Nuclear Antigen-1 genetic variation and potential autoreactive targets relevant to multiple sclerosis pathogenesis". In: *PLoS ONE* 11.2. ISSN: 19326203. DOI: 10.1371/journal.pone.0147567.

Tso, Ken Kai-Yuen et al. (2013). "Complete genomic sequence of Epstein-Barr virus in nasopharyngeal carcinoma cell line C666-1." In: *Infectious agents and cancer* 8.1, p. 29. ISSN: 1750-9378. DOI: 10.1186/1750-9378-8-29.

Tsurumi, T, Daikoku, and Y Nishiyama (1994). "Further characterization of the interaction between the Epstein-Barr virus DNA polymerase catalytic subunit and its accessory subunit with regard to the 3'-to-5' exonucleolytic activity and stability of initiation complex at primer terminus." In: *Journal of virology* 68.5, 3354–3363.

Tsurumi, T et al. (1993). "Functional expression and characterization of the Epstein-Barr virus DNA polymerase catalytic subunit." In: *Journal of virology* 67.8, pp. 4651–8. ISSN: 0022-538X.

Turčanová, Vanda and Per Höllsberg (2004). "Sustained CD8+ T-Cell Immune Response to a Novel Immunodominant HLA-B*0702-Associated Epitope Derived from an Epstein-Barr Virus Helicase-Primase-Associated Protein". In: *Journal of Medical Virology* 72.4, pp. 635–645. ISSN: 01466615. DOI: 10.1002/jmv.20023.

Turk, Susan M et al. (2006). "Antibodies to gp350/220 Enhance the Ability of Epstein-Barr Virus To Infect Epithelial Cells". In: *JOURNAL OF VIROLOGY* 80.19, pp. 9628–9633. ISSN: 0022-538X. DOI: 10.1128/JVI.00622-06.

Tzellos, Stelios and Paul Farrell (2012). "Epstein-Barr Virus Sequence Variation—Biology and Disease". In: *Pathogens* 1.2, pp. 156–174. ISSN: 2076-0817. DOI: 10.3390/pathogens1020156

Tzellos, Stelios et al. (2014). "A Single Amino Acid in EBNA-2 Determines Superior B Lymphoblastoid Cell Line Growth Maintenance by Epstein-Barr Virus Type 1 EBNA-2." In: *Journal of virology* 88.16, pp. 8743–53. ISSN: 1098-5514. DOI: 10.1128/JVI.01000-14.

Uniprot Database. http://www.uniprot.org, accessed last April 2017.

Uniprot Database (accessed last April 2017). http://www.uniprot.org/uniprot/P01106.

Vakiani, E et al. (2008). "Genetic and phenotypic analysis of B-cell post-transplant lymphoproliferative disorders provides insights into disease biology". In: *Hematol Oncol* 26.4, pp. 199–211. DOI: 10.1002/hon.859.

van der Velden, W J F M et al. (2013). "Reduced PTLD-related mortality in patients experiencing EBV infection following allo-SCT after the introduction of a protocol incorporating pre-emptive rituximab." In: *Bone marrow transplantation* 48.11, pp. 1465–71. ISSN: 1476-5365. DOI: 10.1038/bmt.2013.84.

van Diemen, Ferdy R. et al. (2016). "CRISPR/Cas9-Mediated Genome Editing of Herpesviruses Limits Productive and Latent Infections". In: *PLoS Pathogens* 12.6. ISSN: 15537374. DOI: 10.1371/journal.ppat.1005701.

van Gent, Michiel et al. (2014). "Epstein-Barr Virus Large Tegument Protein BPLF1 Contributes to Innate Immune Evasion through Interference with Toll-Like Receptor Signaling". In: *PLoS Pathogens* 10.2. ISSN: 15537374. DOI: 10.1371/journal.ppat.1003960.

van Noort, Johannes M. et al. (2000). *Mistaken self, a novel model that links microbial infections with myelin-directed autoimmunity in multiple sclerosis*. DOI: 10.1016/S0165-5728(00)00181-8.

van Rees, Bastiaan P et al. (2002). "Different pattern of allelic loss in Epstein-Barr virus-positive gastric cancer with emphasis on the p53 tumor suppressor pathway." In: *The American journal of pathology* 161.4, pp. 1207–1213. ISSN: 0002-9440. DOI: 10.1016/S0002-9440(10)64397-0.

Vereide, D T et al. (2014). "Epstein-Barr virus maintains lymphomas via its miRNAs." In: *Oncogene* 33.10, pp. 1258–64. ISSN: 1476-5594. DOI: 10.1038/onc.2013.71. arXiv: NIHMS150003.

Vita, R. et al. (2014). "The immune epitope database (IEDB) 3.0". In: *Nucleic Acids Research* 43.D1, pp. D405–D412. ISSN: 0305-1048. DOI: 10.1093/nar/gku938.

Vockerodt, Martina et al. (2015). "The Epstein-Barr virus and the pathogenesis of lymphoma." In: *The Journal of pathology* 235, pp. 312–22. ISSN: 1096-9896. DOI: 10.1002/path.4459.

Wandinger, K et al. (2000). "Association between clinical disease activity and Epstein-Barr virus reactivation in MS." In: *Neurology* 55.July, pp. 178–184. ISSN: 0028-3878. DOI: 10.1212/WNL.55.2.178.

Wang, Jianbin and Stephen R Quake (2014). "RNA-guided endonuclease provides a therapeutic strategy to cure latent herpesviridae infection." In: *Proceedings of the National Academy of Sciences of the United States of America* 111.36, pp. 13157–62. ISSN: 1091-6490. DOI: 10.1073/pnas.1410785111.

Wang, Peng et al. (2008). "A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach". In: *PLoS Computational Biology* 4.4. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000048.

Wang, Peng et al. (2010a). "Peptide binding predictions for HLA DR, DP and DQ molecules." In: *BMC bioinformatics* 11.1, p. 568. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-568.

Wang, Shanshan et al. (2016). "Identification and Characterization of Epstein-Barr Virus Genomes in Lung Carcinoma Biopsy Samples by Next-Generation Sequencing Technology." In: *Scientific reports* 6, p. 26156. ISSN: 2045-2322. DOI: 10.1038/srep26156.

Wang, Xingang et al. (2010b). "Widespread sequence variation in the Epstein-Barr virus latent membrane protein 2A gene among northern Chinese isolates". In: *Journal of General Virology* 91.10, pp. 2564–2573. ISSN: 00221317. DOI: 10.1099/vir.0.021881-0.

Wang, Yun et al. (2010c). "New variations of Epstein-Barr virus-encoded small RNA genes in nasopharyngeal carcinomas, gastric carcinomas, and healthy donors in northern China". In: *Journal of Medical Virology* 82.5, pp. 829–836. ISSN: 01466615. DOI: 10.1002/jmv.21714.

Wang, Yun et al. (2010d). "Variations of Epstein-Barr virus nuclear antigen 1 gene in gastric carcinomas and nasopharyngeal carcinomas from Northern China". In: *Virus Research* 147.2, pp. 258–264. ISSN: 01681702. DOI: 10.1016/j.virusres.2009.11.010.

Wasil, Laura R. et al. (2013). "The Effect of Epstein-Barr Virus Latent Membrane Protein 2 Expression on the Kinetics of Early B Cell Infection". In: *PLoS ONE* 8.1. ISSN: 19326203. DOI: 10.1371/journal.pone.0054010.

Watanabe, Takahiro et al. (2015). "The Epstein-Barr virus BRRF2 gene product is involved in viral progeny production". In: *Virology* 484, pp. 33–40. ISSN: 10960341. DOI: 10.1016/j.virol.2015.05.010.

Wei, Kuang Rong et al. (2011). "Histopathological classification of nasopharyngeal carcinoma". In: *Asian Pacific Journal of Cancer Prevention* 12.5, pp. 1141–1147. ISSN: 15137368.

Wei, William I. and Jonathan S T Sham (2005). "Nasopharyngeal carcinoma". In: *Lancet*. Vol. 365. 9476, pp. 2041–2054. ISBN: 1474-547X (Electronic)\n0140-6736 (Linking). DOI: 10.1016/S0140-6736(05)66698-6.

White, Robert E. et al. (2012). "EBNA3B-deficient EBV promotes B cell lymphomagenesis in humanized mice and is found in human tumors". In: *Journal of Clinical Investigation* 122.4, pp. 1487–1502. ISSN: 00219738. DOI: 10.1172/JCI58092.

WHO (2014). *Nasopharyngeal Carcinoma. 2014 Review of Cancer Medicines on the WHO List of Essential Medicines*. Tech. rep.

Wilkinson, Dianna E. and Sandra K. Weller (2004). "Recruitment of Cellular Recombination and Repair Proteins to Sites of Herpes Simplex Virus Type 1 DNA Replication Is Dependent on the Composition of Viral Proteins within Prereplicative Sites and Correlates with the Induction of the DNA Damage Response". In: *Journal of Virology* 78.9, pp. 4783–4796. DOI: 10.1128/JVI.78.9.4783-4796.2004. eprint: http://jvi.asm.org/content/78/9/4783.full.pdf+html.

Witter, R L et al. (1970). "Isolation from turkeys of a cell-associated herpesvirus antigenically related to Marek's disease virus". In: *American Journal of Veterinary Research* 31, pp. 525–538.

Woisetschlaeger, M et al. (1990). "Promoter switching in Epstein-Barr virus during the initial stages of infection of B lymphocytes." In: *Proceedings of the National Academy of Sciences of the United States of America* 87.5, pp. 1725–9. ISSN: 0027-8424.

Wu, Guocai et al. (2012). "Characterization of Epstein-Barr virus type 1 nuclear antigen 3C sequence patterns of nasopharyngeal and gastric carcinomas in northern China".

In: *Archives of Virology* 157, pp. 845–853. ISSN: 03048608. DOI: 10.1007/s00705-012-1241-y.

Wucherpfennig, Kai W. and Jack L. Strominger (1995). "Molecular mimicry in T cell-mediated autoimmunity: Viral peptides activate human T cell clones specific for myelin basic protein". In: *Cell* 80.5, pp. 695–705. ISSN: 00928674. DOI: 10.1016/0092-8674(95)90348-8.

Wykes, Michelle N. et al. (2014). "Malaria drives T cells to exhaustion". In: 5.MAY. ISSN: 1664302X. DOI: 10.3389/fmicb.2014.00249.

Xia, Tianli et al. (2008). "EBV microRNAs in primary lymphomas and targeting of CXCL-11 by ebv-mir-BHRF1-3". In: *Cancer Research* 68.5, pp. 1436–1442. ISSN: 00085472. DOI: 10.1158/0008-5472.CAN-07-5126.

Xiao, Jianqiao et al. (2008). "The Epstein-Barr virus BMRF-2 protein facilitates virus attachment to oral epithelial cells". In: *Virology* 370.2, pp. 430–442. ISSN: 00426822. DOI: 10.1016/j.virol.2007.09.012.

Yajima, Misako, Teru Kanda, and Kenzo Takada (2005). "Critical Role of Epstein-Barr Virus (EBV)-Encoded RNA in Efficient EBV-Induced B-Lymphocyte Growth Transformation". In: *Journal of Virology* 79.7, pp. 4298–4307. DOI: 10.1128/JVI.79.7.4298-4307.2005. eprint: http://jvi.asm.org/content/79/7/4298.full.pdf+html.

Yamamoto, Takenobu and Keiji Iwatsuki (2012). "Diversity of Epstein-Barr virus BamHI-A rightward transcripts and their expression patterns in lytic and latent infections". In: *Journal of Medical Microbiology* 61.PART 10, pp. 1445–1453. ISSN: 00222615. DOI: 10.1099/jmm.0.044727-0.

Yang, Ying et al. (2014). "Sequence Analysis of EBV Immediate-Early Gene BZLF1 and BRLF1 in Lymphomas". In: January. DOI: 10.1002/jmv.

Yang, Ziheng (2007). "PAML 4: Phylogenetic analysis by maximum likelihood". In: *Molecular Biology and Evolution* 24.Ml, pp. 1586–1591. ISSN: 07374038. DOI: 10.1093/molbev/msm088.

Yang, Ziheng and Willie J Swanson (2002). "Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes." In: *Molecular biology and evolution* 19, pp. 49–57. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a003981.

Yao, Q Y et al. (1996). "Frequency of multiple Epstein-Barr virus infections in T-cell-immunocompromised individuals". In: *J Virol* 70.8, pp. 4884–4894. ISSN: 0022-538X.

Yao, Q Y et al. (1998). "Epidemiology of infection with Epstein-Barr virus types 1 and 2: lessons from the study of a T-cell-immunocompromised hemophilic cohort". In: *Journal of Virology* 72.5, pp. 4352–4363. ISSN: 0022-538X.

Yates, J L, S M Camiolo, and J M Bashaw (2000). "The minimal replicator of Epstein-Barr virus oriP." In: *Journal of virology* 74.10, pp. 4512–4522. ISSN: 0022-538X. DOI: 10.1128/JVI.74.10.4512-4522.2000.

Yates, J L et al. (1996). "Comparison of the EBNA1 proteins of Epstein-Barr virus and herpesvirus papio in sequence and function." In: *Virology* 222.1, pp. 1–13. ISSN: 0042-6822. DOI: 10.1006/viro.1996.0392.

Yeh, Te-Shien et al. (1997). "ESequence Variations Between Two Epstein-Barr Virus LMP 1 Variants Have No Effect on the Activation of NF-kappa-B Activity". In: *DNA and Cell Biology* 16.11, pp. 1311–1319.

Yin, Yili, Bénédicte Manoury, and Robin Fåhraeus (2003). "Self-inhibition of synthesis and antigen presentation by Epstein-Barr virus-encoded EBNA1." In: *Science (New York, N.Y.)* 301.5638, pp. 1371–4. ISSN: 1095-9203. DOI: 10.1126/science.1088902.

Young, L S et al. (1987). "New type B isolates of Epstein-Barr virus from Burkitt's lymphoma and from normal individuals in endemic areas." In: *The Journal of general virology* 68 (Pt 11, pp. 2853–2862. ISSN: 0022-1317 (Print). DOI: 10.1099/0022-1317-68-11-2853.

Young, Lawrence S, John R Arrand, and Paul G Murray (2007). "Chapter 27 EBV gene expression and regulation". In: *Human Herpesviruses*. Ed. by Ann Arvin et al. Cambridge: Cambridge University Press. Chap. EBV gene expression and regulation.

Young, Lawrence S. and Christopher W. Dawson (2014). *Epstein-Barr virus and nasopharyngeal carcinoma*. DOI: 10.5732/cjc.014.10197.

Young, Lawrence S, Lee Fah Yap, and Paul G Murray (2016). "Epstein – Barr virus : more than 50 years old and still providing surprises". In: *Nature Publishing Group* 16.12, pp. 789–802. ISSN: 1474-175X. DOI: 10.1038/nrc.2016.92.

Yuen, Kit-san et al. (2015). "CRISPR/Cas9-mediated genome editing of Epstein-Barr virus in human cells." In: *The Journal of general virology* 96.Pt 3, pp. 626–36. ISSN: 1465-2099. DOI: 10.1099/jgv.0.000012.

Yuen, Kit-San et al. (2017). "Suppression of Epstein-Barr virus {DNA} load in latently infected nasopharyngeal carcinoma cells by CRISPR/Cas9". In: *Virus Research*, pp. –. ISSN: 0168-1702. DOI: https://doi.org/10.1016/j.virusres.2017.04.019.

Zeng, Mu-sheng et al. (2005). "Genomic Sequence Analysis of Epstein-Barr Virus Strain GD1 from a Nasopharyngeal Carcinoma Patient †". In: *Society* 79.24, pp. 15323–15330. DOI: 10.1128/JVI.79.24.15323.

Zhang, Xiao-Shi et al. (2002). "The 30-bp deletion variant: a polymorphism of latent membrane protein 1 prevalent in endemic and non-endemic areas of nasopharyngeal carcinomas in China". In: *Cancer Letters* 176.1, pp. 65 –73. ISSN: 0304-3835. DOI: http://dx.doi.org/10.1016/S0304-3835(01)00733-9.

Zhang, Xiao Shi et al. (2004). "V-val subtype of Epstein-Barr virus nuclear antigen 1 preferentially exists in biopsies of nasopharyngeal carcinoma". In: *Cancer Letters* 211.1, pp. 11–18. ISSN: 03043835. DOI: 10.1016/j.canlet.2004.01.035.

Zhao, Bo et al. (2011). "Epstein-Barr virus nuclear antigen 3C regulated genes in lymphoblastoid cell lines." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.1, pp. 337–42. ISSN: 1091-6490. DOI: 10.1073/pnas.1017419108.

Zhou, X G et al. (2001). "Epstein-Barr virus gene polymorphisms in Chinese Hodgkin's disease cases and healthy donors: identification of three distinct virus variants." In: *The Journal of general virology* 82.Pt 5, pp. 1157–1167. ISSN: 0022-1317.

Zimber, Ursula et al. (1986). "Geographical prevalence of two types of Epstein-Barr virus". In: *Virology* 154.1, pp. 56–66. ISSN: 10960341. DOI: 10.1016/0042-6822(86)90429-0.

Zimmermann, J and W Hammerschmidt (1995). "Structure and role of the terminal repeats of Epstein-Barr virus in processing and packaging of virion DNA." In: *Journal of Virology* 69.5, pp. 3147–55. eprint: http://jvi.asm.org/content/69/5/3147.full.pdf+html.

zur Hausen, A et al. (2004). "Epstein-Barr virus in gastric carcinomas and gastric stump carcinomas: a late event in gastric carcinogenesis". In: *Journal of Clinical Pathology* 57.5, pp. 487–491. ISSN: 0021-9746. DOI: 10.1136/jcp.2003.014068.

zur Hausen, Axel et al. (2000). "Unique transcription pattern of Epstein-Barr virus (EBV) in EBV-carrying gastric adenocarcinomas: Expression of the transforming BARF1 gene". In: *Cancer Research* 60.10, pp. 2745–2748. ISSN: 00085472. DOI: http://dx.doi.org/10.1016/S0016-5085(00)82313-6.