

# **Genetic determinants of EBV infection in Lymphoblastoid Cell Lines**

*Witold Czyz*

*Saint Cross College*

*University of Oxford*

**Hilary Term 2014**

**A thesis submitted in partial fulfilment of the requirements for  
the degree of Doctor of Philosophy at the University of Oxford.**



## Abstract

### Genetic determinants of EBV infection in Lymphoblastoid Cell Lines

Epstein-Barr Virus (EBV), a ubiquitous herpesvirus that infects over 95% of the adult human population, has been implicated in the aetiology of a range of autoimmune diseases and tumours. In some of these disorders such as post-transplant B-cell lymphomas, EBV acts as a direct causal factor, in others, like Hodgkin's disease and nasopharyngeal carcinoma, it is an important co-factor. Additionally, EBV infection has been linked to several other diseases, most notably Multiple Sclerosis through positive correlation with the occurrence of Infectious Mononucleosis – a benign lymphoproliferative disease caused by primary EBV infection. The key feature of most EBV-disease associations is the ability of the virus to infect and transform human B- T- NK- and epithelial cells using a set of transcripts and proteins, some of which act as oncogenes. While it is evident that EBV viral load and gene expression may be correlated with the course of disease or even directly contributing to its pathology, the genetic determinants of EBV uptake, expression and its proliferative capacity remain unresolved.

This project aimed to investigate the genetic determinants of EBV copy number and EBV latency gene expression for human B-cells immortalised by EBV *in vitro* and transformed into permanently growing lymphoblastoid cell lines (LCLs), as a model for early-stage EBV infection in naïve B-cells. LCL samples studied have been sourced from several different populations, the HapMap Project, the 1000 Genomes Project as well as British MRC-A family cohort. Methods used encompass quantification of viral expression and copy number using TaqMan and SybrGreen PCR techniques, followed by statistical association tests conducted using Plink, Merlin and MatrixEQTL. EBV QTLs identified by the assays were next subjected to a meta-analysis in GWAMA. Two most significant eQTLs were also selected for a replication experiment in an independent panel of newly generated LCLs and validated in peripheral blood B-cells sourced from the same donors.

Multiple significant and suggestive expression and copy number QTLs were identified. However, most of these associations have not been replicated in more than a single cohort. The relatively small sample size of most cohorts tested as well as population structure posed

a limitation. Some findings merit attention, particularly the presence of statistically significant viral eQTLs within or close to *CSMD1* locus in two different cohorts, and finding of a significant EBV eQTL in a SNP associated with type 1 diabetes risk and located close to *IL2RA*, an immune-response gene harbouring multiple autoimmune disease risk loci. Suggestive associations were also identified in the 1000 Genomes Project samples by the copy number assay which resulted in the most robust test conducted. These encompassed an association to the *PRDM9* locus as well as to a gene involved in TGF- $\beta$  secretion. This is particularly interesting since TGF- $\beta$  signal promotes lytic replication in EBV-infected B-cells and a consistent significant correlation between EBV lytic expression and increased viral copy number has been identified.

In conclusion, although no significant association has been consistently replicated, the project provided several suggestive EBV QTL candidates with plausible biological links to EBV infection and replication, which could be studied further in independent experiments.

## **Acknowledgements**

I would like express my deepest gratitude to my supervisor, Dr Julian Knight. This project would never be possible without his help and guidance. It has been a privilege to work in this most kind environment.

I am much indebted to Dr Albertus Mohr, for the inspiration to conduct the EBV QTL research as well as help and advice at its onset.

I would like to thank my collaborators, Professor William Cookson, Professor Miriam Moffat, Professor Liming Liang et al. 2013. Permitting me to use the biological materials from the MRC-A cohort as well as providing me with the genotype and the expression data enabled the project to expand and formed a solid foundation for my thesis. Thank you for the assistance with technical problems concerning Merlin and the family-based statistical analysis. Thanks to Kenny Wong for preparing all the RNA samples for shipment and to Narelle Maugeri for permitting me to use some of her LCL cDNA material.

I am most grateful to Dr Vivek Naranbhai for his optimism, kind help and valuable advice, especially for the meta-analysis of my results, which could have not been accomplished without his involvement.

Also to Dr Peter Humburg, for explaining the use of MatrixEQTL as well as solving many of the bioinformatic issues I encountered, and to Dr Jean Baptiste-Cazier for providing help and advice on the use of PLINK. I would like to give thanks to my fellow students Katharina Plant, Evelyn Lau and Seiko Machino for introducing me to the lab and molecular biology techniques, and patience in answering my countless technical questions. Finally, special thanks to Kat and Christine Blancher for taking the time to read and check a substantial part of my thesis; and to Kat and Evelyn for helping me with the Biobank samples.

## Contents

<b>Chapter 1. Introduction</b>	<b>1</b>
1.1 Overview	1
1.2 EBV Biology, genome and latency	1
1.3 Mechanism of EBV Infection	5
1.3.1 Introduction	5
1.3.2 Viral entry	7
1.3.3 Latency III expression	8
1.3.4 Transformation, growth and latent viral replication	15
1.3.5 Germinal centre maturation, differentiation and latency II	20
1.3.6 Viral entry in the immune memory pool and latency I and 0	22
1.3.7 Global expression reprogramming during EBV infection	23
1.4. EBV-associated diseases	24
1.4.1. Background	24
1.4.2 Disorders of the lymphatic tissue	26
1.4.2.1 Introduction	26
1.4.2.2 Primary Infection and Infectious Mononucleosis	26
1.4.2.3 Chronic active EBV infection	29
1.4.2.4 Lymphoproliferative disorders	29
1.4.2.5 Monomorphic lymphomas	31
1.4.2.6 Post-transplant, HIV-associated and senile lymphoproliferative disorder	33
1.4.2.7 X-linked lymphoproliferative disorder (XLP)	34
1.4.2.8 Burkitt's lymphoma	35
1.4.2.9 Hodgkin's disease	38
1.4.3 Non B-cell tumours	40
1.4.3.1 Introduction	40
1.4.3.2 NK/T cell lymphoma	40
1.4.3.3 Nasopharyngeal carcinoma (NPC)	41

1.4.3.4 Gastric carcinoma	43
1.4.4 Other EBV-associated disorders	44
1.4.5 EBV-related autoimmune diseases	44
1.5 Host genetics and EBV Infection	47
1.6 Genetic determinants of gene expression	49
1.6.1 Transcript level as mediator of genetic effects on the phenotype	49
1.6.2 Mapping gene expression as a quantitative trait	51
1.6.3 Genetic mapping and GWAS	52
1.6.4 eQTLs in post-GWAS analysis	55
1.6.5 Studies of human and model organism eQTLs	56
1.7. LCLs as a model of EBV-infection	57
1.8. Previous work and rationale for the thesis	60
1.8.1 LMP1 effects on human expression in hypoxia study	60
1.8.2 Rationale	61
1.8.3 Hypothesis	63
1.9. Project's aims	64
<b>Chapter 2. Materials and Methods</b>	<b>67</b>
2.1 Samples	67
2.1.1. HapMap	68
2.1.2. MRC-A	68
2.1.3 1000 Genomes	69
2.1.4 Namalwa LCL	69
2.1.5 Oxford Biobank	70
2.2 RNA work (Oxford Biobank and MRC-A samples)	70
2.2.1 RNA Isolation	70
2.2.2 Reverse Transcription (RT) PCR	71
2.3 Quantitative PCR (qPCR)	71
2.3.1. SYBR Green Quantitative Real Time PCR	72

2.3.2 TaqMan Quantitative Real Time PCR	73
2.4 Housekeeper Assay	74
2.4.1 BestKeeper	74
2.4.2 geNorm	76
2.5. eQTL Mapping	76
2.5.1 Plink	77
2.5.2 Merlin	78
2.5.3 Matrix eQTL	79
2.6 eQTL data integration	80
2.7 GWAMA	80
2.8. Oxford Biobank Whole Blood Sample preparation	82
2.8.1 Sample processing	82
2.8.2. B-cell sorting	82
2.8.3. Stimulation and harvesting of B cells	83
<b>Chapter 3. Mapping EBV transcript QTLs</b>	<b>84</b>
3.1 Introduction	84
3.2 Aims of the chapter	84
3.3 Characterisation of cis effects of the candidate LMP1 eQTL, rs1913243	86
3.4 EBV latency eQTLs in the Hypoxia study LCLs	90
3.4.1. BART	92
3.4.2 EBER1	92
3.4.3. EBER2 and EBNA1	95
3.4.4. EBNA2	96
3.4.5. EBNA3 family	98
3.4.6 LMP1	101
3.4.7 LMP2	103
3.4.8 EBNA-LP104	
3.5 EBV latency eQTLs in the MRC-A panel	104

3.5.1 Housekeeper assay	105
3.5.2 Latency eQTL assay results	106
3.5.3 eQTLs for both EBV and human transcripts	118
3.6 EBV eQTLs from RNA-seq experiments	119
3.6.1 RNA-seq EBV latency and lytic eQTLs	120
3.7 Discussion	128
3.7.1. LMP1A eQTL	128
3.7.2 Latency eQTLs in the hypoxia study panel	128
3.7.3 Latency eQTLs in the MRC-A panel	130
3.7.4 Latency eQTLs in the RNA-seq panels	133
3.8 Conclusion	137
<b>Chapter 4. EBV copy number QTLs</b>	<b>139</b>
4.1 Introduction	139
4.2 Aims of the chapter	139
4.3 Mapping EBV copy number QTL analysis in HapMap LCLs	140
4.4 EBV copy number QTL analysis in 1000 Genomes LCLs	143
4.5 Discussion	151
4.5.1 Copy number QTLs in HapMap LCLs	151
4.5.2 Copy number QTLs in 1000 Genomes samples	151
4.6 Conclusion	155
<b>Chapter 5. Meta-analysis</b>	<b>156</b>
5.1. Introduction	156
5.2 Aims of the chapter	156
5.3 Results	157
5.3.1 Latency eQTL meta-analysis	157
5.3.2 EBV copy number meta-analysis	170
5.3.3 EBV transcript / copy number correlation	173
5.4 Discussion	175

5.4.1 Latency eQTLs	175
5.4.2 EBV transcript and copy number correlation	176
5.4.3 EBV transcript and copy number correlation	176
5.5 Conclusion	178
<b>Chapter 6. MRC-A latency eQTL follow up</b>	<b>179</b>
6.1 Introduction	179
6.2 Aims of the chapter	179
6.3 Results	180
6.3.1. Study design	180
6.3.2. rs17158630, candidate eQTL for EBNA1	182
6.3.3. rs17339199, candidate eQTL for EBNA3A	185
6.4 Discussion	188
6.5 Conclusion	188
<b>Chapter 7. General Discussion</b>	<b>190</b>
7.1 Overview of the results	190
7.2 Future Research	194
Reference	197
Appendix	220

## **Declaration**

I declare that unless otherwise stated all work presented in this thesis is my own. cDNA and RNA from the Mohr study cohort was extracted and kindly provided by Dr Albertus Jacobus Mohr. RNA, genotypes and global expression data for the MRC-A cohort was provided by Professor William Cookson and Professor Liming Liang. Dr Katharine Plant and Miss Evelyn Lau assisted with aspects of RNA and DNA extraction from the Oxford Biobank samples, cell sorting and B-cell stimulation. The GWAMA meta-analysis was performed with guidance from Dr Vivek Naranbhai.

## **Abbreviations**

AIDS – Acquired Immunodeficiency Syndrome

BL – Burkitt’s Lymphoma

BCR – B-cell Receptor

CNS – Central nervous system

CTL – Cytotoxic T-lymphocytes

CIS – Clinically Isolated Syndrome

ChIP – Chromatin immunoprecipitation

DS – Dyad symmetry

DDR – DNA damage response

EBV – Epstein-Barr Virus

EMSA - Electrophoretic mobility shift assay

eQTL – Expression quantitative trait locus

FAIRE – Formaldehyde Assisted Isolation of Regulatory Elements technique

FR – Family of repeats

GC – Gastric carcinoma

GCen – Germinal centre

GWAS – Genome wide association study

HRS – Hodgkin and Reed–Sternberg (cells)

HD – Hodgkin’s disease

IL – Interleukin, for eg. IL-4, IL-12

IM – Infectious mononucleosis

LCL – Lymphoblastoid cell line

LD – Lymphoproliferative disorder

MAF – Minor allele frequency

MS – Multiple sclerosis

NPC – Nasopharyngeal carcinoma

ORF – Open reading frame

OriP – Origin of plasmid

PBMC – Peripheral blood mononuclear cells

PML – Promyelocytic leukaemia

PTLD – Post-transplant lymphoproliferative disorder

QTL – Quantitative trait loci

RA – Rheumatoid arthritis

QTL – Quantitative trait locus

SLE – Systemic lupus erythematosus

TES – Transcription end site

TR – Terminal repeats

TSS – Transcription start site

WHO – World health organisation

YRI – Yoruba

## **List of Manufacturers**

Bio-Rad; Bio-rad Laboratories Ltd., Hemel Hempstead, Herts UK

Coriell; Coriell Cell Repositories Camden, New Jersey 081303, USA

Invitrogen; Invitrogen, Leek, The Netherlands

Life Technologies; Life Technologies Ltd., Paisly, UK

Miltenyi; Miltenyi Biotec GmbH, Bergisch Gladbach, Germany

PHE – Public Health England, Porton Down, Wiltshire, UK

Qiagen; Qiagen Ltd., Dorking, Surrey, UK

Thermo Scientific; Thermo Fischer Scientific, Massachusetts, USA

## **1. Introduction**

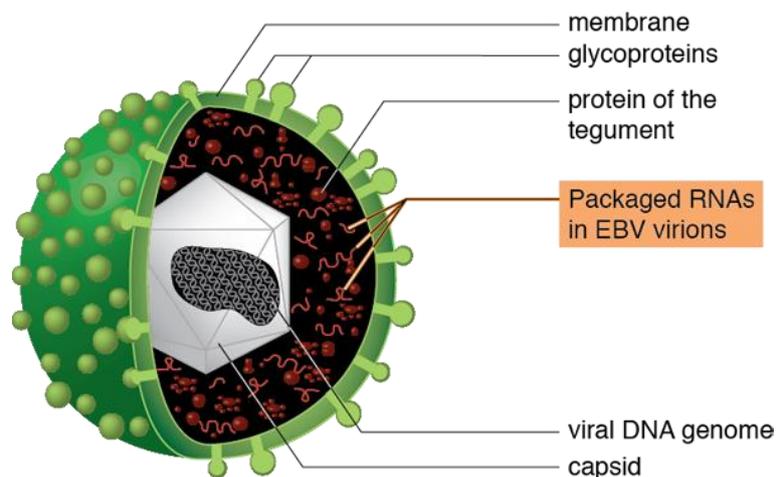
### **1.1 Overview**

There is growing evidence supporting the role of Epstein-Barr Virus (EBV) as an aetiological factor contributing to human cancer and autoimmune disease. Current literature has provided detailed information on viral biology highlighting strong, often causal, relationships between EBV latency proteins, copy number and human disease (Crawford 2001, Hiraki et al. 2001, Hsu and Glaser 2000, Hohaus et al. 2011, Hadinoto et al. 2008). There is also evidence of genetically determined heterogeneity in EBV latent infection, and EBV related disease (Caliskan et al. 2011, Rubicz et al. 2013). This emphasises the importance of genetic factors that may alter viral latent expression and persistence and, consequently, be relevant to EBV-associated disease. This introduction gives a broad background to EBV biology, aiming to review mechanisms for viral infection with a particular focus on latency transcripts and copy number in the context of human cellular pathways and disease association. Evidence implicating viral transcripts and proteins in pathogenesis is presented, and known human genetic regulatory effects on EBV expression and copy number described. A discussion of methods for mapping genetic determinants of transcript expression and the utility of lymphoblastoid cell lines (LCLs) for modelling EBV infection follows. Finally, the rationale, the hypothesis as well as specific aims of the thesis are presented.

### **1.2 EBV Biology, genome and latency**

Since its discovery in 1964, EBV is a member of the Herpesvirus family and has been linked to several different human diseases, in particular a range of lymphoma and carcinoma tumours, but also certain autoimmune diseases like Systemic lupus erythematosus (SLE) and

Multiple Sclerosis (MS) (Hiraki et al. 2001, Crawford 2001, Young and Rickinson 2004). Herpes viruses characteristically establish a lifelong latent infection in their host (Crawford 2001), a process involving viral latency transcripts and their products, in case of EBV, most notably the oncogenic protein LMP1. Latency proteins have thus a direct role in the process of cellular transformation and immortalisation. Their molecular properties enable them, under specific circumstances, to contribute to tumour cell survival and growth.

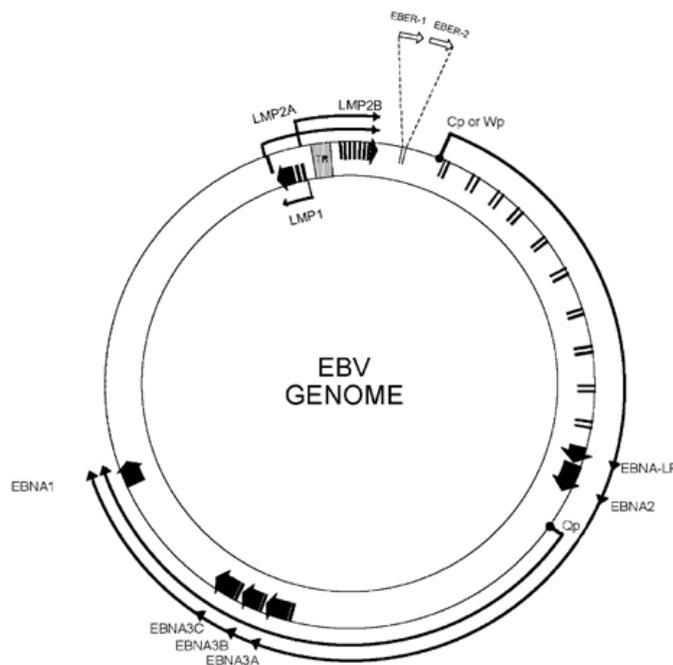


**Figure 1** – Diagrammatic representation of EBV structure (reproduced Jochum et al. 2012). EBV virions consist of three main elements. The lipid envelope membrane, with its glycoproteins which bind B-cell receptors, merges with B-cell membrane and passes the capsid-bound viral DNA into the cytoplasm along with tegument RNAs and proteins that disable intrinsic cellular defences (Tsai et al. 2011). The capsid then travels to nuclear pores and passes the naked EBV DNA into the nucleus.

The EBV genome is 184kb in length comprising 89 genes that are divided into 43 core genes, common to all herpesviruses, and 46 non-core genes of which 28 are EBV-specific (Crawford 2001, Straus et al. 1993, Kieff et al. 2010, Calderwood et al. 2007). The viral genome is contained within an icosahedral nucleocapsid containing linear double stranded DNA molecule and surrounded by an envelope (Figure 1) (Hiraki et al. 2001, Hsu and Glaser 2000).

EBV, which primarily infects B-lymphocytes, hijacks a natural B-cell maturation and differentiation pathway. It thus enters the immune system's long-term memory pool within infected memory B-cells. It can achieve this via 9 key latency proteins and over 11 latency

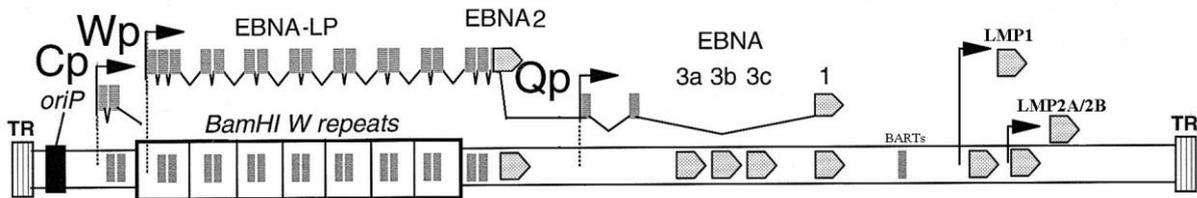
transcripts. Throughout infection, latency transcripts are expressed according to three key latency modes which dramatically alter cellular transcription (Hiraki et al. 2001, Dimmock et al. 2007, Kieff et al. 2010). The latency transcripts include six nuclear protein precursors, EBNA-1, -2, -3A, -3B, -3C and EBNA leader protein (or EBNA-LP), three latent membrane proteins, LMP1, -2A, -2B, and two types of non-translated RNAs called EBV-encoded small RNA 1 and 2 (or EBER-1 and EBER-2) (schematic arrangement of latency genes, Figure 2-3) (Hiraki et al. 2001, Crawford 2001).



**Figure 2** – Circular schematic diagram of EBV Genome (reproduced from Young and Rickinson 2004). Within the B-cell’s nucleus EBV genome forms circular episomes through its terminal repeats (TR – marked in grey on the diagram), and expresses a limited set of latency transcripts, which can be grouped into EBV nuclear antigens (EBNAs), latency membrane proteins (LMPs) and EBV-encoded small RNAs (EBERs). Most EBNAs are expressed from a common polycistronic transcript originating near the Wp or Cp promoter. This transcript contains a highly repetitive leader sequence (which is translated into the EBNA Leader Protein or EBNA-LP). Because each repeat contains a Wp promoter, the length of the leader sequence may vary. EBNA1 can additionally be expressed from its own Qp promoter. LMP and EBERs have their own independent promoters.

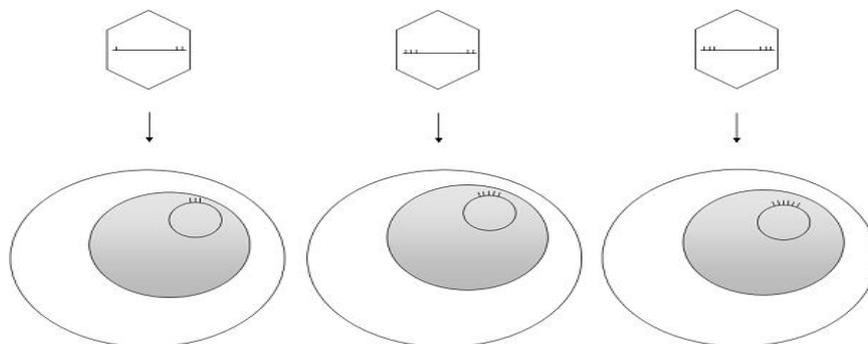
In addition, up to 20 non-translated transcripts from the BamHI region, located between the EBNA and LMP open reading frames (ORFs) are usually expressed (Kieff et al. 2010). These are microRNAs of 22-24 nucleotides and likely form part of the RISC complex modulating

both host and viral mRNA degradation (Swaminathan 2010). Latency proteins are essential for the virus to transform B-cells into continuously proliferating lymphoblasts *in vivo*, or LCLs *in vitro* (Crawford 2001). EBV was first isolated from lymphoma tumour cells highlighting this potentially oncogenic property (Crawford 2001).



**Figure 3** – Schematic EBV linear genome (adapted from Paulson and Speck 1999). EBV is approximately 184 kb long, starting with the TR and origin of plasmid (OriP, necessary for circularisation). Latency transcript ORFs and intronic sequences are indicated by grey rectangles and arrows. Viral promoters (Cp, Wp and Qp, LMP1p and LMP2p) are indicated by black arrows. The ability of each promoter to initiate transcription of a particular set of latency genes is marked by lines joining the introns.

The EBV genome contains repetitive elements and indels, which are the main source of its genetic variation (Tzello and Farrell 2012). The most frequent type of difference is the number of multiple terminal repeats (TR) (Figure 4), which are important for herpesvirus episome’s circularisation and thus necessary for replication (Repic et al. 2010, Collins et al. 2002).



**Figure 4** - EBV terminal repeat variability (reproduced from Takacs et al. 2009). Number of terminal repeats (marked by vertical bars on the figure) is the most frequent source of genomic diversity in EBV. Within B-cell nucleus TRs are bound by EBNA1 and fuse together to form an episome. EBV clonality in infected B-cells is determined by the size of the BamHI fragment containing the variable number of TRs.

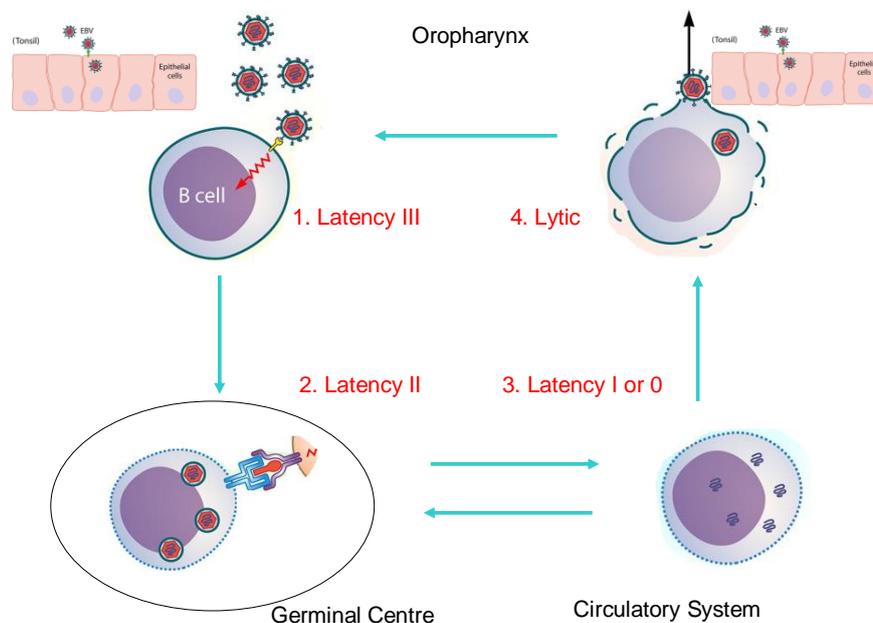
There are two main strains of the virus, distinguished mainly by the sequence of the key viral latency transactivator, EBNA2, and termed EBV A and B, or EBV I and II (Kieff et al. 2010). EBV B or II is found frequently in Africa, but is otherwise rare, while most Western EBV isolates belong to type A. EBV A is characterised by a more efficient B-cell infection and transformation in vitro (Kieff et al. 2010). The EBV isolate B95-8, which has been fully sequenced and used to establish efficient lytic replication in cotton-top tamarin (*Saguinus Oedipus*) lymphocytes for subsequent human B-cell transfection, belongs to strain A (Farrell et al. 1997). Sequence differences in viral strains can also occur in other latency genes like EBNA1, LMPs as well as lytic genes independently of EBNA2 polymorphisms though no phenotypic or disease associations have been reported to these variations (Kieff et al. 2010).

### **1.3 Mechanism of EBV Infection**

#### **1.3.1 Introduction**

During primary infection EBV utilises a limited set of transcripts to establish permanent latent infection. They can be divided into non-translated RNAs, EBV nuclear antigen family proteins (EBNAs) and latency membrane proteins (LMPs). All latency genes play a role in B-cell immortalisation and maintenance of LCLs, however only two, EBNA-2 and LMP1, often termed EBV oncogenes, appear to be absolutely essential for in-vitro immortalisation (Hiraki et al. 2001, Young and Rickinson 2004), although some authors claim the EBNA-3 family and EBNA-LP are also essential (Dawson et al. 2012, Kieff et al. 2010, White et al. 2010, McClellan et al. 2012). Depending on the environment, EBV infection can be divided into 3 stages characterised by distinct patterns of expression called the latency types (Figure 5).

Each stage corresponds to a particular event of the natural B-cell maturation pathway that the virus interferes with, and exploits to establish permanent latency. It is important to differentiate between these stages and expression modes since each has a different aim and is characterised by a unique network of viral-human interactions and thus can give rise to distinct kinds of tumours. The key mediators of viral infection are the EBNA and LMP proteins that interact with multiple human TFs and target several core signalling pathways within the B-cell. These pathways, normally controlling cellular growth and differentiation, are central to viral transformation.

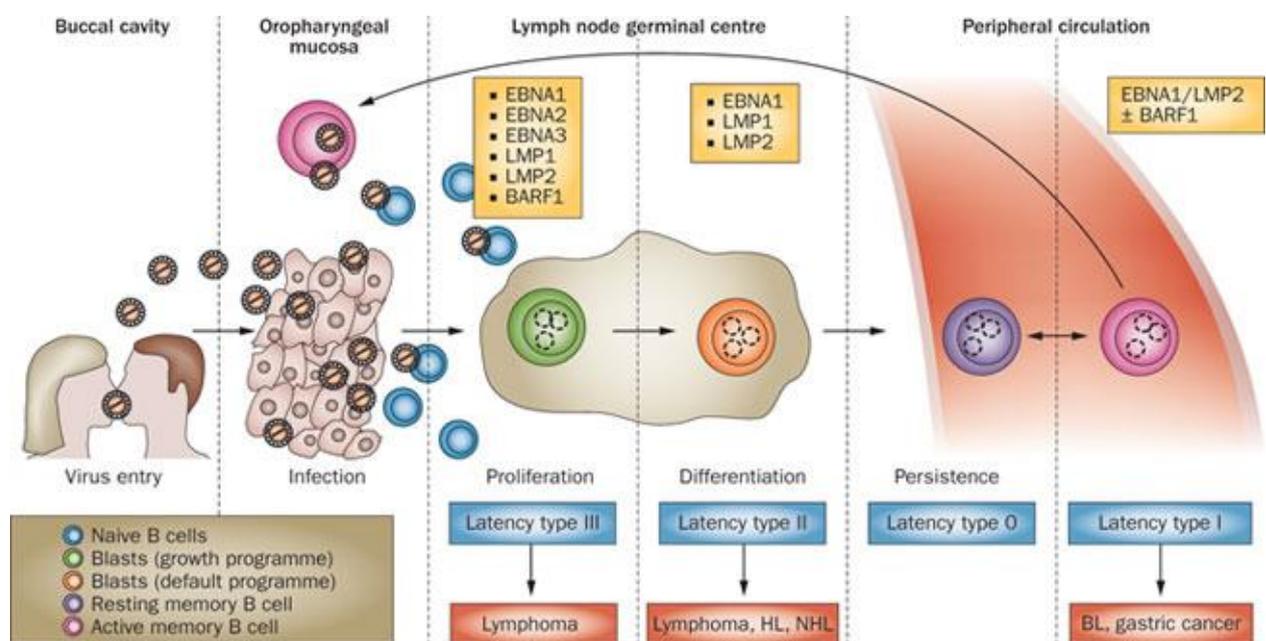


**Figure 5** – EBV Life cycle (adapted from Odumade et al. 2011).

- i) EBV virion particles transmitted in saliva penetrate the porous epithelium of the tonsillar crypts within oropharynx and infect naïve B-cells expressing the full set of latency transcripts (corresponding to stage 1 “Latency III”)
- ii) Activated B-cells travel to lymph nodes within tonsils and form germinal centres in which expression is changed to Latency type II allowing for survival and maturation (stage 2 “Latency II”)
- iii) Mature resting memory B cells exist the germinal centre and enter circulation expressing no or a limited set of latency transcripts (stage 3 “Latency I or 0”)
- iv) Memory B-cells persist throughout the lifetime of the host, re-entering the lymph node and dividing or differentiating into plasma cells at which stage the virus re-activated and switches to lytic expression/replication mode (stage 4 “Lytic mode”)

### 1.3.2 Viral entry

It has been difficult to establish the exact site of viral replication *in vivo* during initial stages of the disease (Crawford 2001, Dimmock et al. 2007). There has been debate over which type of cells constitute the primary target of the EBV infection (Crawford 2001, Dimmock et al. 2007). Current evidence suggests that the virus, transmitted in saliva, travels to the tonsillar crypts in the oropharynx where it infects naïve B-lymphocytes either directly, where the epithelial lining becomes discontinuous, or after an initial round of replication within the epithelium (Hiraki et al. 2001, Crawford 2001, Straus et al. 1993) (Figure 6).



**Figure 6** – EBV Infection (adapted from Bollard et al. 2012 et al). Stage 1 - Viral Entry. EBV is transmitted in saliva. The onset of infection occurs in the oropharynx, in the epithelium of tonsillar crypts or, alternatively, directly at the site of tonsils within a layer of naïve B-cells termed follicular mantle.

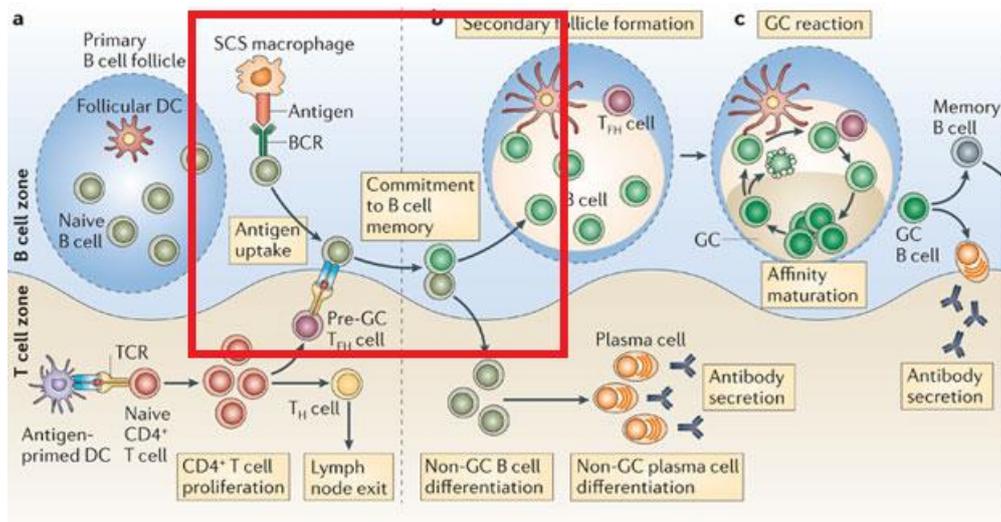
EBV can infect both B-lymphocytes and epithelial cells *in vitro*, though the efficiency and the outcome differs between these tissues (Hiraki et al. 2001, Crawford 2001, Young and

Rickinson 2004). The infection is mediated by the EBV envelope glycoproteins, gp350, which bind to the CD21 receptor on the B-cell surface (Hiraki et al. 2001) and gp42 which binds to the HLA class II co-receptors (Young and Rickinson 2004). Epithelial cells lack these EBV receptors and therefore it is very difficult to infect them *in vitro* (Tsao et al. 2012). However, EBV is able to infect and replicates lytically in epithelial cells in acute infectious mononucleosis (IM) and in nasopharyngeal carcinoma (NPC), so it is possible that under certain specific conditions it can utilise other receptors (Tsao et al. 2012, Young and Rickinson 2004, Buettner et al. 2012).

Viral particles bind host surface receptors and protective capsids, each harbouring a linear copy of the EBV genome, travel through the membrane into the cytoplasm. Through an interaction between the capsid and cell nuclear pores, the viral DNA is transferred into the nucleus. There, viral proteins derived from the tegument facilitate circularisation and chromatinisation of EBV episome (Lieberman 2013).

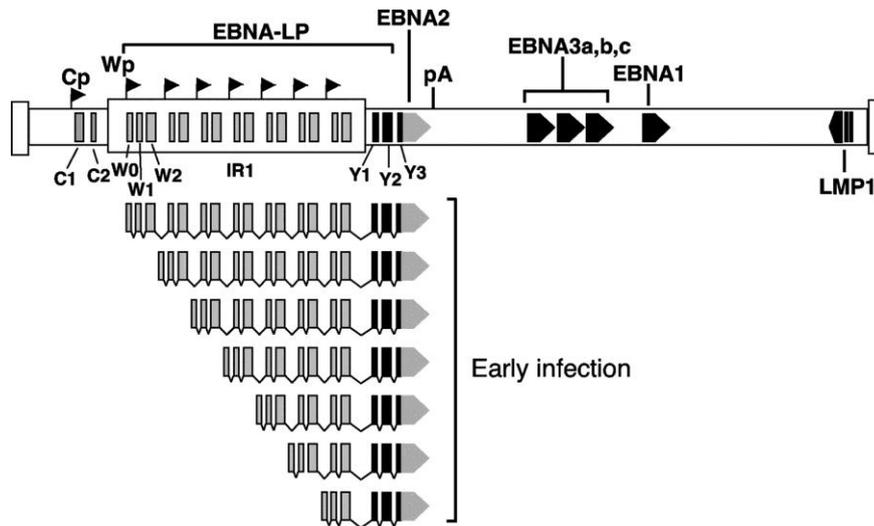
### **1.3.3 Latency III expression**

This stage takes place at the beginning of the infection and it emulates natural B-cell activation by a cognate antigen presented by an antigen presenting cell (for instance a macrophage within a primary follicle), subsequent migration to T/B-cell zone interface and helper T-cell co-stimulation (Figure 7). The aim is to drive proliferation of B-blasts, just as it occurs prior to or shortly after the migration of activated B-cells into a germinal centre (GC) (Pereira et al. 2010). Such proliferation might however under special circumstances evade the immune system's control and turn into a lymphoproliferative disease and eventually a malignant monomorphic lymphoma. Increased number of cellular divisions could also increase the chances of oncogenic mutations (Crawford 2001),



**Figure 7** – Latency type III expression mimics the events which occur naturally at the lymph node – namely the activation of naïve B-lymphocytes by their cognate antigen is replaced by the proliferation signal delivered by EBNA2, and the early pre-germinal centre T-cell help is replaced by the survival signal from the LMP1 oncogene (McHeyzer-Williams et al. 2012).

To drive proliferation EBV uses Latency III mode of expression. Shortly after the viral episome enters into the nucleus, the host RNA polymerase type II initiates transcription of a common EBNA-2/EBNA-LP bi-cistronic transcript from the EBV Wp promoter (Ling 2010, Tierney et al. 1994, Ling et al. 1994, Price et al. 2012, Tempera et al. 2010). Wp is present in multiple copies within each repeat of the major internal repeat called IR-1 (also called Bam HI W repeats), which also encodes the repetitive and variable in length EBNA-LP protein (Figure 8) (Elliott et al. 2004).

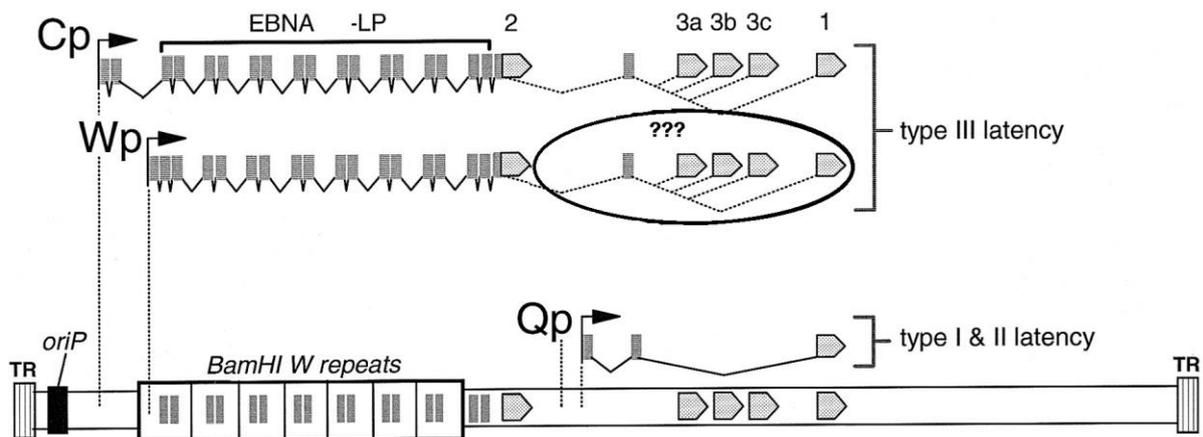


**Figure 8** – Early EBV latency transcripts (Ling 2010). The diagram depicts the linear EBV genome and the variability of early transcript length. Early bicistronic transcripts are transcribed from the EBNA-LP and EBNA2 ORF, and initiated from the Wp promoter. Their leader sequence therefore lacks the C1 C2 introns, characteristic of Cp promoter transcripts. The Wp transcript can be initiated from one of multiple Wp promoters within the highly repetitive IR1 region, as each repeat carries a copy of a Wp promoter. Consequently the size of early transcripts varies, although each carries a copy of EBNA2 at the 3' end.

EBNA-2 is the main viral transactivator and is functionally related to Notch (Young and Rickinson 2004). It is expressed together with EBNA-LP during the earliest stage of infection (Figure 8) and interacts with the Jkappa-recombination-binding protein (RBP-Jkappa, also known as CBF1), found in the Notch signalling pathway (Kato et al. 2011, Kempkes 2010, Rooney et al. 1989). Together with RBP-Jkappa, EBNA-2 acts by decreasing transcriptional repression and activates the expression of a cascade of cellular genes plus other viral latency genes. These include the EBNA3 family, which act as EBNA2 competitors and repressors, EBNA1 as well as the viral membrane proteins LMPs. This occurs by shifting viral expression from the Wp to the Cp promoter. The shift is directly mediated by accumulating EBNA-2, which acts in concert with CBF1 and CBF2 to bind and stimulate the Cp promoter

(Kempkes 2010, Chau et al. 2006). Following the promoter switch, a long polycystronic mRNA template is produced containing all EBNAs.

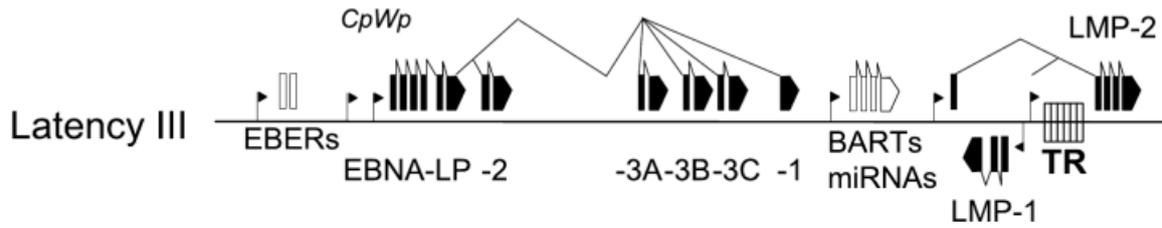
Alternative splicing allows all EBNA family proteins to be synthesized from this single transcript (Lieberman 2013, Evans et al. 1996, Santak 2004). The Wp promoter activity is decreased most likely due to a combination of CpG methylation and competition from the more efficient Cp promoter (Elliott et al. 2004). The Cp promoter is also proposed to be modulated by other TFs, such as SP1, Egr-1 and NF-Y (Kempkes 2010), and binding sites for these factors are found further from the promoter. According to some authors, the EBNA-3 family and EBNA-1 can also be expressed from the Wp promoter at the same time and in addition to the Cp transcripts (Figure 9) (Elliott et al. 2004, Kelly et al. 2006, Young and Rickinson 2004).



**Figure 9** – EBV latency promoters and corresponding transcripts (Paulson and Speck 1999). At the very early stage of infection EBNA2 and EBNA-LP are the first latency transcripts to be expressed. Their expression is initiated from the Wp promoter and once EBNA2 accumulates and transactivates the Cp promoter, transcription of all EBNAs is initiated from Cp promoter. A single polycystronic transcript is produced. There is however evidence that such polycystronic transcripts are also initiated from the Wp promoter (question marks) and present along with Cp-initiated mRNAs in LCLs. There is also evidence that EBNA1 can be independently expressed from its own Qp promoter, this explains the occurrence of EBNA1 in Burkitt lymphoma cells and other cell types, in which it is the only EBV latency protein/transcript present.

EBNA-2 also transactivates LMP-1/LMP2B and LMP2A transcription via RBP-Jk from additional independent promoters, LMP1Ap (or ED-L1) and LMP2Ap (or TP1), which are located near the TR region (Repic et al. 2010, Wu et al. 2000 et al 2000, Takacs et al. 2009). LMP1 can also be expressed irrespective of EBNA2: either when ED-L1 is bound by cellular TFs like STATs, IRF7 and activating transcription factor 4 (ATF4), or from yet another promoter located within the TR, called L1-TR, which is active in nasopharyngeal carcinoma and Hodgkin's lymphoma cells (Repic et al. 2010).

Some authors propose that in order to enable the promoter shift (Table 1) and full latency III expression (Figure 10) the linear genome must form a covalently closed episome via its TRs and therefore EBNA1, which mediates the process, has to be expressed early (Santak 2004, Repic et al. 2010, Lieberman 2013). In this case EBNA1 would be initially expressed from its own independent Qp promoter, before the production of the long Cp transcript (Lieberman 2013). EBNA1 expression is characteristic of all modes, latent or lytic (Frappier 2010). This DNA-binding protein binds two elements within the viral origin of replication (OriP) called the family of repeats (FR) and the dyad symmetry (DS) element, circularising the viral DNA into an episome (Frappier 2010). Circularisation of viral DNA may be the crucial pre-condition allowing for its replication in both lytic (rolling circle replication) and latent (bidirectional replication) state, tethering to human chromosomes, packaging and maintenance into daughter host cells (Frappier 2010). Circularisation is mediated by EBNA1 bound to OriP which acts as a platform for recruiting additional host cellular machinery including the origin of replication complex (ORC), the minichromosome maintenance complex (MCM), shelterin components and polymerase II (Young and Rickinson 2004, McFadden and Luftig 2013).



**Figure 10** – Full latency III expression (Kieff et al. 2010). Untranslated RNAs are marked by hollow rectangles. There are 12 latency transcripts expressed at Latency III.. These include two EBER small RNAs, seven EBNAs, and three LMPs -LMP1, LMP2A and LMP2B. LMP2 function is conferred by the LMP2A protein. There is also LMP2B whose function is poorly understood and which may act to inhibit LMP2A. Both have their own independent promoters and are translated from different transcripts, however their ORFs overlap and differ only by one intron which is unique to LMP2A. Additionally a variable set of up to twenty BART miRNAs is expressed in Latency III.

In the model proposed above, EBNA1 together with CTCF mediate the promoter shift by forming DNA loops which join Cp promoter to viral OriP on one side, and OriP to the LMP promoter on the other. An enhancer sequence called FR within OriP acts then as a universal enhancer for all latency transcripts (Lieberman 2013, Kempkes 2010). EBNA1 can not only enable the promoter shift and the transcription of other EBV latency genes, but also acts as an independent global transcriptional activator for all latency genes when bound to the FR enhancer sequence within oriP. It could also regulate latent transcription through its interaction with nucleosome assembly protein 1 (Frappier 2010).

Latency	Cp	Qp	Wp	EBER1p	EBER2p	LMP1+2p	BARF0p	BARF1p
<b>“Cp on”</b>								
<b>type III</b>	+	-	+/-	+	+	+	+	-
<b>“Cp off”</b>								
<b>type II</b>	-	+	-	+	+	+/-	+	+
<b>type I</b>	-	+	+/-	+	+	-	+	-
<b>type 0</b>	-	+/-	-	?	?	-	?	?

**Table-1**(adapted from Takacs et al. 2009) – depicts the activity of promoters (listed in the top panel) in different types of latency (listed in the side panel). EBV promoters can be broadly grouped into two categories – those controlling the expression of the EBNA family of proteins and those controlling the expression of LMP transmembrane proteins as well as untranslated viral RNAs, namely EBERs (EBER1 and EBER2) and BARTs (BARF0 and BARF1). Latency types can also be grouped depending on promoter usage. Particularly latency type III differs from others since it relies on the Cp promoter for the transcription of all EBNA.

Once within the cell, EBV forms a circular episome, which is likely to subvert cellular double stranded damage response mechanism and evade degradation (Lieberman 2013, McFadden and Luftig 2013). Still, only a minority of genomes manage to undergo successful circularisation, which requires DNA processing, homologous recombination and ligation. The majority of infections are halted by the cell defense mechanisms, and only 1% of infected B-cells progress into viral latency (Lieberman 2013). Once circularised, the viral genome may continue into an intermittent stage of rolling circle replication, typical of lytic proliferation, before key early latency transcripts are expressed, or can very occasionally integrate into the cellular genome (Lieberman 2013). If this occurs, EBV frequently loses its virulence (Lieberman 2013). Within the nucleus the viral tegument protein BNRF1 mediates recruitment of histones and chromatinisation of the EBV episome by interacting with

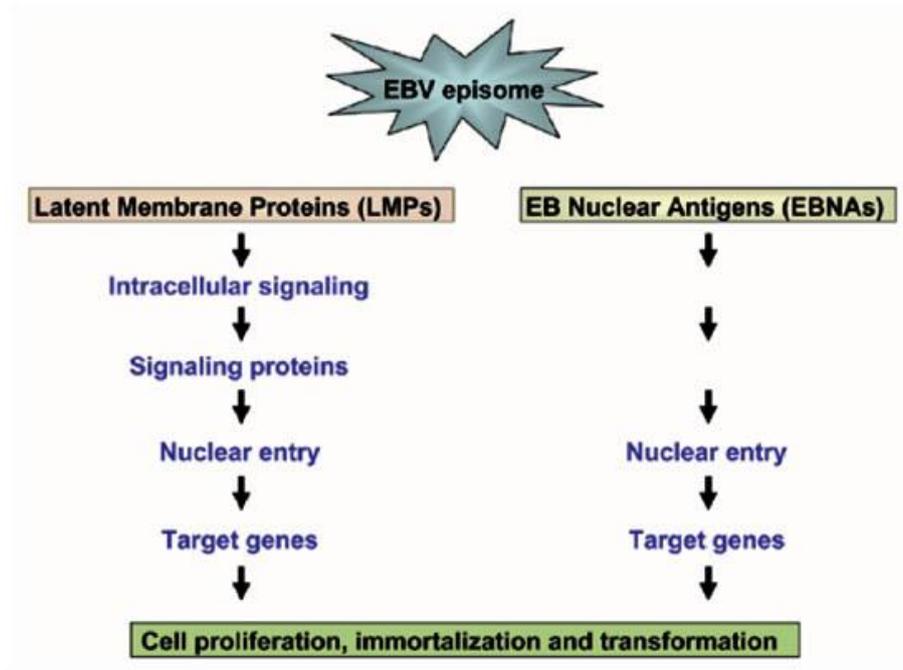
promyelocytic leukaemia nuclear bodies and preventing epigenetic silencing of viral repetitive elements and transcription repression (Lieberman 2013).

EBNA1 may also have other functions and could modulate specific host gene and protein activity directly, including elements of the p53 pathway that leads to speculation about its function in tumorigenesis (Frappier 2010). However, it is difficult to prove the regulatory effects of EBNA1 in isolation from its general enhancer effect on the other latency proteins (Frappier 2010). EBNA1 expression has been implicated in lytic infection of epithelial cells through displacement of promyelocytic leukemia (PML) nuclear bodies - proteins responsible for preventing viral replication (Frappier 2010, Frappier 2011, Frappier 2012, McFadden and Luftig 2013).

Once the promotor shift occurs, a full set of 12 latency transcripts is expressed. This type of latent expression is called latency III and is also characteristic of circulating B-cells in IM (Calderwood et al. 2007). It causes an antigen-like activation and rapid proliferation of B-cells. Normally, naïve B cells are activated by antigen binding B-cell receptor (BCR) with helper T cell co-stimulation and then migrate to proliferate within germinal centres (Kuppers 2009). In viral infection, the activation step is replaced by latency III expression whose components target multiple signalling pathways (Thorley-Lawson 2001).

#### **1.3.4 Transformation, growth and latent viral replication**

Latency III expression aims to stimulate cell's growth (Figure 11). Two cellular pathways play crucial role in B-cell immortalisation: the Notch pathway and NF- $\kappa$ B pathway.



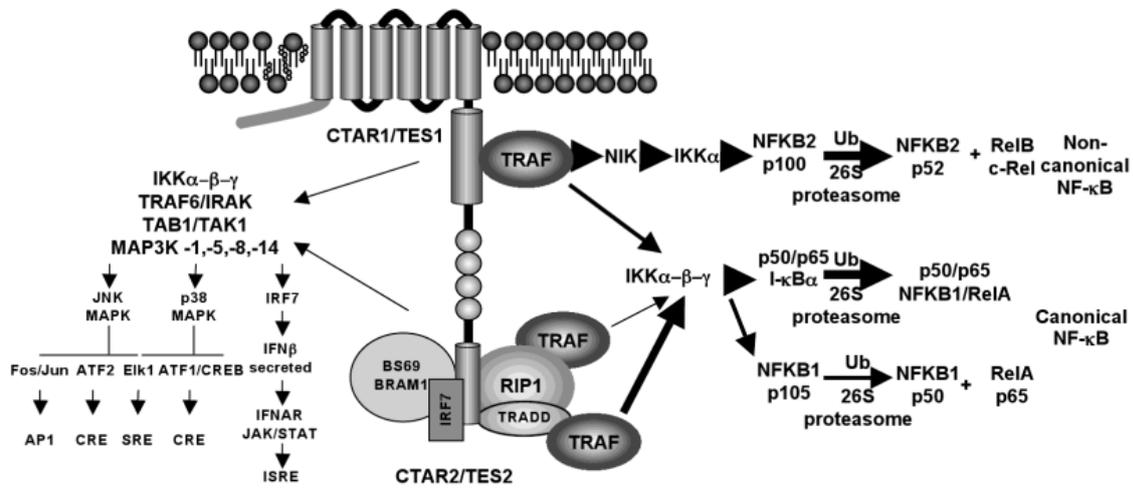
**Figure - 11** (Liu et al. 2006) - Viral proteins in B-cell transformation process. The two families of EBV latency proteins, with the central function of EBNA2 and LMP1, target intrinsic cellular pathways in the cytoplasm (intracellular signalling of LMPs). They affect gene expression in the nucleus either directly (EBNA2-RBPJk) or indirectly through downstream components of the NF- $\kappa$ B pathway like p50 and p65. Their ultimate aim is to upregulate B-cell proliferation and maturation.

Through interacting with RBP-Jkappa, EBNA-2 upregulates the downstream targets of the Notch-signalling pathway, CD21 CD23 and c-myc promoters, thus driving proliferation and preventing differentiation (Hiraki et al. 2001, Straus et al. 1993). In normal B-cells Notch acts as a tumour suppressor and inhibits B-cell maturation (Kempkes 2010). However, the response to Notch varies depending on the signal intensity and developmental stage of the cell, and active Notch may act synergistically with CD40 and BCR signalling delivered by LMP1 and LMP2 to stimulate proliferation (Kempkes 2010) during the initial stage of the infection, but it has to be finally reduced in order to enable B-cell differentiation and latency (Hardie 2010). This is likely to be why EBNA-2 expression is switched off later, at the germinal centre maturation stage (Thorley-Lawson 2001, Young and Rickinson 2004).

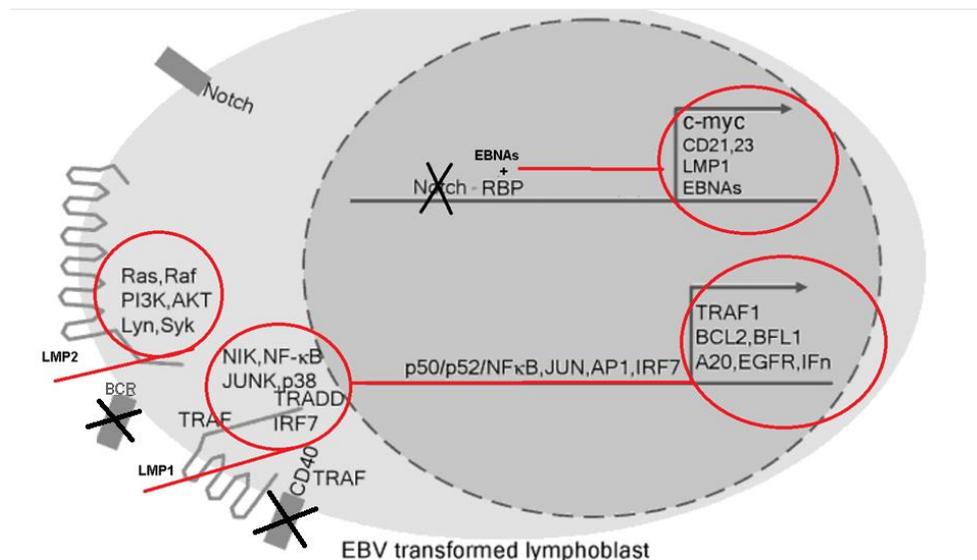
Cells with active Notch and c-myc are normally prone to apoptosis, but this is likely prevented by the EBNA3 family which silences the expression of the pro-apoptotic Bcl2-like protein 11 by recruiting chromatin-silencing histone deacetylase complex (HDAC) (Thorley-Lawson and Allday 2008, Ocheni et al. 2010). EBNA-LP upregulates EBNA-2 driven expression via chromatin modifiers, mainly by relocalising transcription repressors like HDAC4/5 away from EBNA-2 target promoters, but also by displacing Sp100A, a cellular cofactor associated with chromatin activation, from PML nuclear body proteins. PML nuclear body proteins are responsible for an intrinsic defence mechanism against viral infection and normally silence viral expression (Ling 2010). In contrast to EBNA-LP, EBNA3 proteins reduce the EBNA2 driven activation by competing with and displacing EBNA2 from RBP-Jkappa (Sims et al. 2010). EBNA-3 proteins and EBNA-LP are crucial for the fine-tuning of viral expression (Young and Rickinson 2004). EBNA3C is also involved in abrogation of the innate DNA damage response by interacting with host TFs, which would otherwise stop the growth of EBV infected B-cells by cell cycle arrest (McFadden and Luftig 2013).

LMPs are the other group of proteins produced at full latency III. Recombinant EBV, genetic and biochemical experiments indicate that LMP-1 is a key viral oncogene essential for B-cell transformation (Kieff et al. 2010, Izumi 2010) (Figures 12 and 13). This transmembrane protein is present in all EBV-positive malignancies and has oncogenic effects on cultured rodent fibroblasts and human epithelial cells (Dawson et al. 2012, Izumi 2010). This is because LMP-1 mimics a constitutively activated tumour necrosis factor receptor (TNFR) eliciting pleiotropic survival effects (Izumi 2010). In particular, it interacts with tumour necrosis factor receptor-associated factors (TRAFs) acting as CD40 and drives nuclear factor- $\kappa$  B (NF- $\kappa$ B) canonical and non-canonical signalling pathway (Izumi 2010, Hiraki et al. 2001, Young and Rickinson 2004). It also activates, through its CTAR/TES intracellular domain, mitogen-activated protein kinase (MAPK) and the downstream cascade of cytokines

including extracellular-regulated protein kinase (ERK), c-Jun NH2-terminal kinase (JNK) and p38 (Hiraki et al. 2001, Thorley-Lawson 2001, Young and Rickinson 2004). Thirdly, the CTAR/TES residues, activate interferon regulatory transcription factor 7 (IRF7) and stimulate the JAK/STAT signalling pathway through interferon-beta secretion (Izumi 2010). The overall effect of LMP-1 on cellular expression is similar to that of a T-helper cell's signalling, and ensures survival of germinal centre B-cells carrying EBV (Thorley-Lawson 2001).



**Figure 12** - LMP-1 target pathways (Izumi 2010). LMP1 mimics CD40 and its CTAR/TES1 intracellular domain associates with TRAFs but also engages TRADDs and RIP to activate canonical and non-canonical NF-κB, the main target pathway, as well as JNK and p38. Additionally, LMP1 also activates JAK/STAT pathway through induction of interferon regulatory factor 7 (IRF7).



**Figure 13** – Human-EBV molecular interactions (Adapted from Kieff et al. 2010). The diagram shows the molecular interactions of the key EBV proteins. LMP1 binds TRAFs and TRADDs in effect substitutes CD40 function in NF- $\kappa$ B signalling. It also activates c-JUN and IRF7. LMP2A substitutes BCR signalling and modulates Ras/MAPK signalling. EBNA2 substitutes intra-nuclear Notch and up-regulates c-myc as well as CD21 and CD23.

The action of LMP-2 resembles BCR signalling and enables infected B-cells to evade immune system check points (Kieff et al. 2010). LMP-2 blocks BCR-associated tyrosine kinase activation preventing apoptosis and viral re-activation (Bieging et al. 2010, Cohen 2000, Dawson et al. 2012). The action of LMP-2 elicits an intact BCR receptor signal activating the Ras-pathway and promoting survival (Thorley-Lawson 2001). EBER RNAs are likely to also have a role in preventing apoptosis (Crawford 2001, Swaminathan 2010).

In combination, the latency transcripts mimic antigen-driven stimulation and T-cell help. The latency modes are summarised in Table 2 (source: Hiraki et al. 2001).

Mode/Transcript	BART	EBER	EBNA-1	EBNA-2	EBNA-3	EBNA-LP	LMP1	LMP2	Typical of EBV-infected cells in:
LATENCY I	+	+	+	-	-	-	-	-/+	BL
LATENCY II	+	+	+	-	-	-	+	+	HD, NPC*, GC*, T/NK cell*
LATENCY III	+	+	+	+	+	+	+	+	IM, LPD
LATENCY 0	+	+	-	-	-	-	-	-	

**Table-2:** EBV Latency modes (information from Hiraki et al. 2001, Crawford 2001) - the table summarises variations in transcript expression at three key stages of EBV infection. Each combination corresponds to one of latency modes discussed in the main text. Presence or absence of a specific transcript is indicated by “+” or “-“. The last field lists diseases or conditions which a particular type of latency is typical of. \*In some diseases, LMP-1 and LMP-2 expression can vary or be heterogenic.

### 1.3.5 Germinal centre maturation, differentiation and latency II

This is the stage of viral infection in which EBV changes its expression to a different type in order to secure its long-term survival. To achieve this it must guide the B-cell through the maturation process that normally takes place within a germinal centre evading all security check-points. This means that EBV can rescue faulty and crippled B-cells that would otherwise undergo apoptosis (Spender and Inman 2011 Amon and Farrell 2005 Mancao et al. 2005). This may be of significance as some malignant tumours cells display a germinal centre B-cell phenotype and EBV expression pattern (Rezk and Weiss 2007, Brink et al. 1997). EBV might thus rescue B-cells harbouring oncogenic mutations which are “arrested” at latency III or II stage and cannot complete the maturation process (Thorley-Lawson 2001).

Activated B-cells migrate into follicles where they multiply and form germinal centres. Within the germinal centre they would normally undergo rounds of clonal expansion, class-switch recombination, selection and affinity maturation. Other cells such as follicular dendritic cells, macrophages and helper T cells would provide co-stimulatory signals and regulate B-cell maturation (Kuppers 2009).

B-cells proliferate in the dark zone of the follicle where somatic hypermutations take place. Somatic hypermutation enables the diversification of the variable (V) region genes of immunoglobulin (Ig) heavy and light chains generating different BCR affinity (Thorley-Lawson 2001, Kuppers 2009). The cells then undergo a check. Autoreactive and inefficient B-cells die through apoptosis or are eliminated by macrophages, while high-affinity cells are positively selected by co-stimulation from dendritic and helper T cells in another follicular compartment called the light zone (Kuppers 2009). Each cell normally undergoes a series of replication and selection before maturing into a memory B-cell (Kuppers 2009). Here, in the lymph nodes, the viral expression pattern changes to Latency II, likely in response to signals present within the follicles (Thorley-Lawson 2001). EBNA-2 -3 and -LP expression is shut down and proliferation stops while LMP1 and LMP2 deliver signals necessary for survival of the B-cells and their differentiation into memory cells (Thorley-Lawson 2001). EBNA-2 expression shut down is obligatory since the protein is the main factor behind EBV-driven proliferation preventing the B-cell from acquiring a germinal centre phenotype because of the constitutive activation of the Notch pathway (Thorley-Lawson and Allday 2008). The exact cause of the latency switch is unknown (Thorley-Lawson 2001). There is evidence that through interacting with RBP-Jkappa they recruit HDACs and the C-terminal-binding protein 1 (CtBP) complex to the Cp promoter facilitating transcriptional repression by histone modifications which repress transcription and silence the main viral latency promoter (Thorley-Lawson and Allday 2008, Sims et al. 2010). As a result, EBNA-3 proteins switch

EBNA-2 expression off altogether stopping proliferation and enabling B cells to differentiate out into memory cells (Thorley-Lawson and Allday 2008, Sims et al. 2010). Consequently, EBV-directed growth may be directed by a negative feedback loop mediated by EBNA-2 and EBNA-3 expression. It is likely that additional signals present in the lymphoid tissue are necessary to tilt the balance between EBNA-2 and EBNA-3 (Thorley-Lawson and Allday 2008). In the germinal centre, presence of interleukin(IL)-4, IL-10 and IL-21 is likely to activate LMP1 and LMP2 expression in the absence of EBNA2 (Thorley-Lawson and Allday 2008).

### **1.3.6 Viral entry in the immune memory pool and latency I and 0**

This is the final stage of EBV latent infection that ends with the virus drastically reducing its activity so that it can evade the immune system. At this stage infected B-cells cannot be detected by CTLs and the virus persists undisturbed. Lymphoid tumours showing the restricted EBV expression typical of latency I and 0 are limited to Burkitt's lymphoma and HIV-associated plasmablastic lymphoma (Rezk and Weiss 2007).

The mature B-cell exits the germinal centre and lymph node, and enters the circulation escaping from interleukin stimulation and other intra-nodal signals. The cell responds to the change by expressing EBNA-1 and viral RNAs (latency type I) or un-translated RNA only (Latency 0) (Crawford 2001, Straus et al. 1993, Thorley-Lawson and Allday 2008). At this stage the Cp promoter becomes methylated, repressing transcription. EBNA-1 expression is directed from the Qp promoter (Lieberman 2013). EBNA1 can be intermittently expressed as it elicits a weak immune response in contrast to other EBV antigens since it is inefficiently degraded and presented by the MHC class I receptors on cell surface (Hiraki et al. 2001, Crawford 2001, Frappier 2010). Its presence is required when the carrier B-cell occasionally

divides (Thorley-Lawson 2001). Some memory B-cells that re-enter the tonsils can then re-express latency type II, which is likely to be important for cell division and their long-term survival (Thorley-Lawson 2001). Occasionally memory B-cells travel back to the lymph nodes and, if activated by an antigen, differentiate terminally into plasma cells. This most likely causes the virus to enter lytic phase and accounts for low-level viral lytic proliferation in the oropharyngeal epithelium and shedding into saliva (Crawford 2001).

### **1.3.7 Global expression reprogramming during EBV infection**

Viral infection and establishment of latency induces significant global changes in gene expression in both host and viral genomes. These changes are induced by CpG methylation, histone and higher order chromatin modifications. Each latency type can be characterised by a distinct epigenotype (Lieberman 2013). By modulating levels of key chromatin modifiers like H4K20me3, H3K27me3, H3K9me3, H4K20me3 and H3K9me3, EBV decreases chromatin silencing and increases promoter accessibility eliciting global activation of thousands of genes, including *HOX* and *ZNF* (Hernando et al. 2014). There is evidence that this effect is EBV-specific and distinct from growth-promoting CD40L/IL-4 stimulation and that it is largely independent of EBNA2 action (Hernando et al. 2014).

## 1.4. EBV-associated diseases

### 1.4.1. Background

Properties of the viral latency transcripts can, under certain specific conditions, cause disease (Crawford 2001). Since its discovery, EBV has been linked to a range of diseases that can broadly be grouped into autoimmune diseases and tumours (Maeda et al. 2006) (Tables 3 and 4). The latter can be further grouped into lymphomas, lymphoproliferative disorders, mesenchymal and epithelial malignancies.

Disease	Prevalence of EBV / disease tissue	Cofactors
X-linked lymphoproliferative disease	Nearly 100%	XLPS mutations
Infectious Mononucleosis	Nearly 100%	
B-cell lymphoproliferative disease	90% post-transplant 50% HIV-associated peripheral lymphoma ~100% HIV-associated primary CNS lymphoma	Immunosuppression
Burkitt's lymphoma	96-100% in Malaria endemic areas 10%-70% in sporadic cases 30%-40% in AIDS	Immunosuppression <i>c-myc</i> deregulation malaria HIV
Hodgkin's lymphoma	40%-80%  (60% in children, 80% in young adults)	IM
T/NK cell lymphoma	10%	Immunosuppression, chronic IM
Primary effusion lymphoma	70%-80%	
Nasopharyngeal carcinoma	~100% non-keratinizing 30%-100% keratinizing (higher in endemic areas)	Genetic factors Environment (diet)
Gastric carcinoma	100% undifferentiated lymphoadenocarcinoma	

	5%-15% adenocarcinoma	
Oral hairy leukoplakia	100%	

**Table-3** – EBV-associated disorders summary (Crawford 2001, Macsween and Crawford 2003) – diseases in which EBV has been proven to play a role of a causal factor or a co-factor are listed below in the first field; second field quotes the prevalence of EBV in disease-affected tissues; third field outlines any additional co-factors promoting the disease.

<b>Disease</b>	<b>EBV association</b>
SLE	Aberrant T-cell response against EBV, Increased seroprevalence, Elevated anti-EBV antibodies (also predating disease onset), Increased EBV load
RA	Aberrant T-cell response against EBV, Elevated anti-EBV antibodies, Increased EBV load
MS	Aberrant T-cell response against EBV, Increased seroprevalence, Elevated anti-EBV antibodies (also predating disease onset), positive association with IM, Increased EBV load (partial evidence)

**Table-4** – EBV-associated autoimmune diseases – table delineates the evidence and nature of EBV association with the three autoimmune diseases – Systemic Lupus Erythomatosus (SLE), Rheumatoid Arthritis (RA) and Multiple Sclerosis (MS).

In most of these conditions the virus acts as an important co-factor. Of particular importance are the multiple interactions and effects of viral oncogene proteins aimed to ensure long term viral persistence and survival; these become tumourigenic in abnormal conditions. As the

viral proteins are dependent on cellular pathways, determining their regulation may indicate potential therapeutic targets.

## **1.4.2 Disorders of the lymphatic tissue**

### **1.4.2.1 Introduction**

Because EBV takes advantage of natural B-cell development in order to infect and persist, it is not surprising that the main and most frequent type of EBV-associated diseases are lymphomas and lymphoproliferative disorders. In some, EBV is the direct cause. This is supported by the finding that many patients with or at risk of EBV-positive lymphomas have higher anti-EBV antibody levels, and in all EBV-positive tumours the infection occurs prior to cancer development (Kieff et al. 2010).

### **1.4.2.2 Primary Infection and Infectious Mononucleosis**

Exposure to EBV is ubiquitous and over 90-95% world's adult population becomes infected in their lifetime (Crawford 2001, Straus et al. 1993, Thorley-Lawson 2001, Hopwood and Crawford 2000). Many individuals will become infected by multiple herpes viruses in their lifetime, and can also be infected by multiple EBV strains (Crawford 2001). During primary infection, EBV stimulates B-cell proliferation and replicates with the host cell by bidirectional replication, if latent. It also replicates lytically (by rolling circle replication) in a subset of 0.1-0.0001% B cells, which accounts for its presence in the saliva of the infected individuals (Crawford 2001, Dimmock et al. 2007). The number of transformed B-cells varies

from individual to individual, but falls within the range of 1 to 50 cells per million (Hiraki et al. 2001, Crawford 2001).

Most infections, particularly in developing countries, occur sub-clinically in early childhood (by the age of 3-5 years), however in developed countries up to 50% individuals may remain uninfected into adolescence and 30% to 50% out of these individuals, upon primary infection, will develop Infectious Mononucleosis (IM) with glandular fever, pharyngitis and adenopathy (Ascherio and Munger 2007, Crawford 2001, Dimmock et al. 2007, Straus et al. 1993, Rubicz et al. 2013). These symptoms are caused by strong humoral and cellular immune responses with substantial CD8+ cytotoxic T lymphocytes (CTLs) activation (Crawford 2001). In acute infection, viral BCRF1 (high homology to interleukin-10) and BARF1 proteins help evade the immune response by inhibiting the synthesis of interferon gamma and alpha, respectively (Hiraki et al. 2001, Straus et al. 1993).

EBV infection and B-cell proliferation is controlled by the host CTLs and between 0.1% to 3% of peripheral blood CD8+ and CD4+ T cells in every individual who has undergone infection in the past are EBV-specific (Kieff et al. 2010). In healthy individuals, CTLs eventually overcome the virus eliminating the replicating B-cells. Once this primary infection is contained, a state of equilibrium follows in which the virus is mostly present in a pool of transformed resting memory B-cells. As the immune system eliminates replicating EBV-infected B-cells (which display lytic antigens easily detectable by CTLs), lytic replication stops almost entirely and shedding of viral particles into saliva decreases. However it does not stop entirely because a small subset of EBV-infected peripheral memory B-cells recirculate periodically back into lymph nodes where, in response to unknown factors, they undergo differentiation into plasma cells which automatically elicits lytic replication (Thorley-Lawson 2001, Straus et al. 1993). Unlike other Herpesviruses, EBV does not re-activate in

healthy individuals and acute infection converts to asymptomatic persistent infection (Straus et al. 1993).

In primary infection, EBV can cause mild fever, and in approximately 30-70% of adolescents and young adults it can lead to IM (Tattevin et al. 2006, Hopwood and Crawford 2000). This type of infection can be seen as a mild lymphoproliferative disease and its characteristic feature is a rapid increase in infected B-lymphocytes (Crawford 2001). The immune system cannot control the infection at its early onset which leads to high levels of lytic replication and subsequent re-infection (Thorley-Lawson 2001, Hadinoto et al. 2008). A positive feedback loop is established allowing EBV to infect up to 25% of memory B-cells (Thorley-Lawson 2001, Hadinoto et al. 2008). This may be caused by a higher viral dose at inoculation or an initial nonspecific and crossreactive T-cell response (Crawford 2001). As more B-cells are transformed and more cytotoxic T cells are committed to fight the virus, a massive inflammatory and polyclonal immune response takes place eliciting characteristic disease symptoms of IM (Tattevin et al. 2006). Older patients in particular are at risk of severe IM that may require hospitalisation and, in extreme cases, even result in death (Tattevin et al. 2006). The course of IM is partly determined by the ability of the virus to enter different cell types, including already differentiated memory B-cells, which once transformed cannot further differentiate and continue to proliferate instead (Thorley-Lawson 2001).

Occasionally IM also manifests in persistent fever, severe hepatosplenomegaly, severe cytopenia, coagulopathy, central nervous system (CNS) abnormality, vascular dysfunction and histiocytic erythrophagocytosis and is recognised clinically as EBV-associated hemophagocytic lymphohistiocytosis (EBV-HLH) (Kasahara et al. 2001). This type of extreme primary EBV infection can often result in premature death. Infection plus activation of T cells observed in this disorder is considered to be the underlying cause of abnormal regulation of cytotoxic T-lymphocyte (CTL) response (Kasahara et al. 2001).

In rare cases, primary acute EBV infection in children is accompanied by dermatosis with confluent papules or papulovesicles present symmetrically on the face and buttocks, and sometimes also with fever, hepatosplenomegaly or lymphadenopathy (Gianotti-Crosti syndrome) (Chisholm and Lopez 2011, Llanora et al. 2012).

#### **1.4.2.3 Chronic active EBV infection**

Normally, viral proliferation is overcome by the immune system and the positive loop of proliferation and re-infection is broken, however very rarely the symptoms persist for 6 months or longer and lead to other complications involving major organs including acute renal failure, hepatic failure, gastrointestinal or pulmonary haemorrhage (Maeda et al. 2006, Tsai et al. 2011). This happens in non-immunocompromised patients, whose T and NK cells become infected by EBV in chronic infection, but can also occasionally occur during normal, acute primary EBV infection (Kasahara et al. 2001, Tsai et al. 2011). In chronic active EBV infection (CAEBV), the mechanism through which EBV manages to establish infection of T or NK cells (or T cells in EBV-HLH) is unknown (Kasahara et al. 2001).

#### **1.4.2.4 Lymphoproliferative disorders**

Any factor altering the balance between the virus-driven low-level lytic replication and the level of cytotoxic T-cells that control infection, can result in uncontrolled B-cell proliferation. This in turn leads to benign hyperplasia or malignancy (Maeda et al. 2006, Muti et al. 2003, Muti et al. 2005). This often happens in immuno-compromised patients and leads to cellular hyperproliferation and a range of EBV-positive lymphoproliferative disorders (Dojcinov et al. 2011, Hsu and Glaser 2000, Dimmock et al. 2007, Crawford 2001, Straus et

al. 1993). This may also explain why higher EBV load is seen in immunoblastic lymphomas of immune-compromised patients (Thorley-Lawson and Allday 2008). EBV positive lymphoproliferative disorders can be classified into several subtypes (Campo et al. 2011, Shroff and Rees 2004). Benign forms include lesions present mostly in tonsils, adenoids or lymph nodes and not affecting other tissues (Ibrahim and Naresh 2012). These lesions are typically found in IM-like hyperplasia (Shroff and Rees 2004). Polymorphic (as all types of B-cell morphology and differentiation stages are present) lymphoproliferative disorder affects extranodal tissues with destruction of tissue structure and infiltrations of T-cells and macrophages (Ibrahmi, Shroff and Rees 2004). B-cells in polymorphic lymphoproliferative disorder may be derived from a single clone (with specific immunoglobulin gene rearrangement), or less typically, they may be polyclonal (though always monoclonal for IgH and the EBV genome) (Ibrahim and Naresh 2012, Chadburn 2013). Early lesions and polymorphic lymphoproliferative disorder are characterised by latency type II and III expression, and can regress upon alleviation of immunosuppression (Shroff and Rees 2004).

<b>World Health Organization Classification of Post-transplant Lymphoproliferative Disorder (PTLD)</b>	
<b>Category</b>	<b>Subtype</b>
Early lesions	Plasmacytic hyperplasia Infectious mononucleosis-like lesion
Polymorphic PTLD Monomorphic PTLD (classify according to lymphoma they resemble)	B-cell neoplasms – Diffuse large B-cell lymphoma – Burkitt lymphoma – Plasma cell myeloma – Plasmacytoma-like lesion – Other <sup>a</sup>  T-cell neoplasms – Peripheral T-cell lymphoma NOS – Hepatosplenic T-cell lymphoma – Other

**Table 5** – Non-malignant lymphoid tumours. Subtypes of PTLD (WTHO classification after Jascobson and Lacasce 2010).

EBV-positive benign lymphoproliferative disorders encompass primary and acquired immunodeficiency lymphoproliferative disorder, senile lymphoproliferative disorder, some AIDS-associated lymphoproliferative disease, B-cell post-transplant lymphoproliferative disease (PTLD, Table 5), lymphoproliferative disorder caused by treatment with methotrexate or tumour necrosis factor- $\alpha$  (TNF- $\alpha$ ) antagonists, and X-linked lymphoproliferative syndrome (Rivat et al. 2013, Hopwood and Crawford 2000, Dojcinov et al. 2011). In all these conditions the disease is a direct result of viral transformation left unchecked. Only in X-linked lymphoproliferative syndrome, a genetic fault disables successful immune response against the virus. Typically, early-onset lesions are polymorphic and polyclonal expressing latency III growth programme, typical of uncontrollably proliferating B-cells. Key viral oncogenes EBNA2 and LMP1 are the main cause of outgrowth (Hopwood and Crawford 2000). However, additional oncogenic mutations will lead to malignant monoclonal tumours, where frequently a more restricted pattern of viral expression is observed (Hopwood and Crawford 2000, Tattevin et al. 2006). This tumorigenesis may be enhanced by increased proliferation that increases the chances of mutation (Crawford 2001).

#### **1.4.2.5 Monomorphic lymphomas**

If untreated, polymorphic forms of diseases like PTLD, AIDS-associated or senile lymphoproliferative disorder, may eventually convert to a monomorphic cancer, usually a diffuse large B-cell lymphoma (Montanari et al. 2010). Monomorphic disorders are malignant and aggressive, and unlike early lesions and polymorphic lymphoproliferative disorders, usually do not respond to reduction of immunosuppression (Ibrahim and Naresh 2012). If unchecked, typically the disorder progresses through the full spectrum from an early

hyperplastic lesion to a malignant monomorphic lymphoma (Chadburn 2013). Most monomorphic lymphomas fall, according to the World Health Organization, into Hodgkin lymphomas, and non-Hodgkin lymphomas that include diffuse large B-cell lymphoma and Burkitt lymphoma - the most common types. However EBV DNA and expression has been associated with other rare types of lymphatic cancer like pyothorax-associated lymphomas, CNS lymphomas in non-immunocompromised patients as well as lymphomatoid granulomatosis (another rare type of non-Hodgkin B-cell lymphoma) (Cohen 2000).

There are also other types of non-Hodgkin's EBV-positive lymphomas, but which affect other immune cells. These fall within the T- and NK-cell lymphoma category (Montanari et al. 2010).

In contrast to polymorphic lymphoproliferative disorder, in which EBV onco-proteins directly drive B-lymphocyte proliferation, in malignant monomorphic lymphomas EBV is instead a co-factor and additional somatic mutation are critical for the malignancy to arise (Ibrahim and Naresh 2012, Chadburn 2013, Thorley-Lawson 2001). Uncontrolled viral infection in immune-compromised people might promote the outgrowth of monomorphic lymphoma through increased rate of B-cell division and limited ability of the immune system to eliminate faulty cells harbouring oncogenic mutations. Evidence for this is seen in tumour progression from hyperplasia, through benign polymorphic disorder into malignancy, as previously discussed. Oncogenic alterations may be sufficient to cause the disease on their own and consequently not all malignant lymphomas are EBV positive (Thorley-Lawson 2001).

#### **1.4.2.6 Post-transplant, HIV-associated and senile lymphoproliferative disorder**

Following a transplant, immune suppression is necessary to avoid graft rejection. Under these conditions, PTLD can occur. Early symptoms usually manifest as an IM-like benign lymphoproliferative illness characteristic of viral reactivation, which occurs in 1-10% transplant patients, particularly those who were previously unexposed to the virus (Hopwood and Crawford 2000). A similar mechanism occurs in a subgroup of B-cell lymphomas affecting HIV carriers. HIV-associated lymphoproliferative disorders are caused by HIV-induced immunosuppression, which abrogates immune system's control of proliferation of EBV-infected B-lymphocytes (Carbone et al. 2008). This causes an IM-like disease which can however convert to an aggressive B-cell lymphoma, the second most common malignancy in AIDS patients after Kaposi sarcoma caused by the very closely related Kaposi sarcoma virus (Roland and Stock 2003, Rezk and Weiss 2007). Latency III expression is typical of PTLD and HIV-associated lymphoproliferative disorder (Carbone et al. 2008).

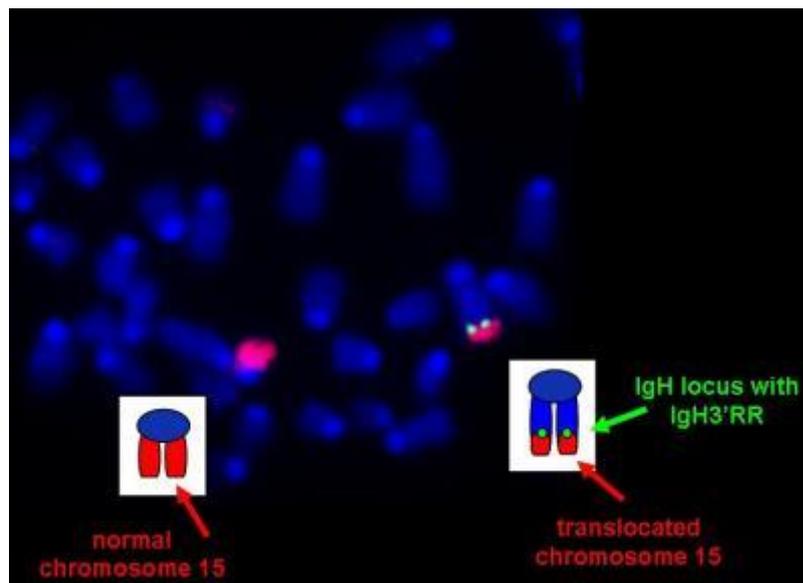
Senile lymphoproliferative disorder with a prevalence of 6-11% in individuals aged 60 and over, was identified in 2003 (Schiozawa et al. 2007). As previously discussed, the phenotype varies between polymorphic polyclonal disease (proposed to be typical of early-onset disease) and monomorphic disease where LMP1 or both LMP1 and EBNA2 are present (type II or type III latency) (Shmiovoyama et al. 2006, Oyama et al. 2003, Gibson and Hsi 2009). It occurs in patients with no apparent immunodeficiency although age-related impaired T cell immunity and immune system deterioration has been proposed as an underlying cause (Mueller et al. 2007, Schiozawa et al. 2007, Dojcinov et al. 2011).

#### **1.4.2.7 X-linked lymphoproliferative disorder (XLP)**

X-linked lymphoproliferative disease (XLP) is genetic heritable disease that manifests in uncontrolled primary Epstein–Barr virus (EBV) infection, lymphoma or dysgammaglobulinaemia (Rezaei et al. 2011). The onset of the disease most often ensues upon primary EBV infection by the age of 2.5 years and patients appear healthy beforehand (Rezaei et al. 2011). Rapid expansion of T and B-cells, most occurs likely caused by T- and natural killer (NK)-cell dysfunction with 100% mortality by the age of 40 (Coffey et al. 1998, Engel et al. 2003, Rezaei et al. 2011). Fatal IM with EBV–associated hemophagocytic lymphohistiocytosis is the cause of death in approximately 50-60% cases of XLP, while 26% cases are caused by malignant monoclonal lymphomas and over 30% by dysgammaglobulinaemia (Marsh et al. 2010, Coffey et al. 1998). Most cases of the disorder are caused by germline mutations to *SH2D1A*, encoding SLAM-associated protein (or SAP – an intracellular adaptor and immunomodulator of NK and T cells) that is expressed in T and NK cells (Rezaei et al. 2011). Mutant SAP, controlling the transduction of signals mediated by the SLAM receptor family and antigen-driven NK and T cell activation, impairs NK and CD8<sup>+</sup> T-cell cytotoxicity towards EBV-transformed B-cells (Filipovich et al. 2010, Rezaei et al. 2011, Rivat et al. 2013). A similar but less frequent form of XLP is caused by mutations of *XBIRC4* encoding apoptosis inhibitor XIAP (Schmid et al. 2011, Marsh et al. 2010). XIAP mutations are likely causing elevated levels of apoptosis among cytotoxic T cells controlling EBV primary infection, leading to a hyperproliferation of transformed B-cells and EBV–associated hemophagocytic in 90% of cases (Marsh et al. 2010). In XIAP XLP, EBV association may be slightly weaker than in the first type (Marsh et al. 2010, Rezaei et al. 2011), however recent studies suggest over 80% of XIAP XLP cases are EBV-positive (Schmid et al. 2011).

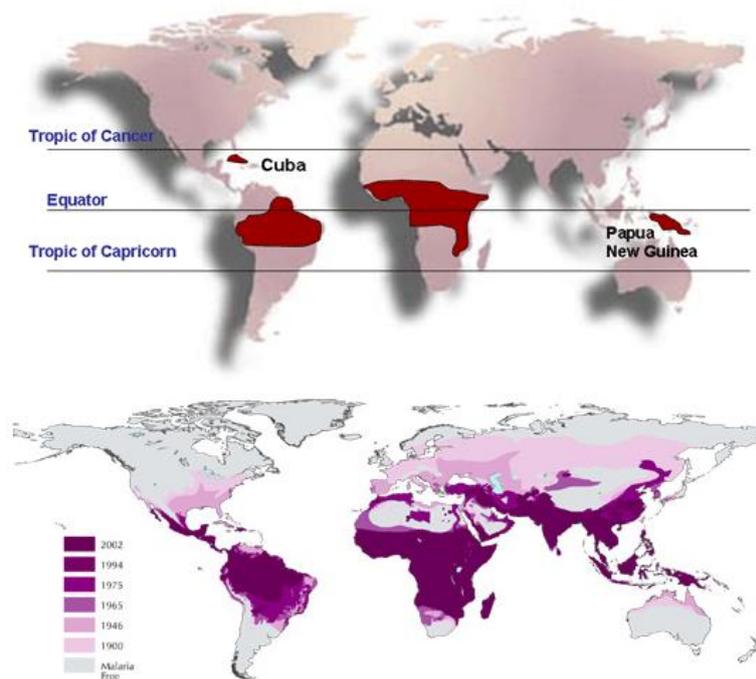
#### 1.4.2.8 Burkitt's lymphoma

EBV itself was first discovered and isolated from cultured cells of Burkitt's lymphoma (BL) (Thorley-Lawson and Allday 2008). BL is a tumour most prevalent in sub-Saharan Africa and coincides strongly with malaria distribution (Thorley-Lawson and Allday 2008). Almost 100% of African BL cases are EBV-positive, however this association is not consistent and in other parts of the world cases can fall below 25% (Thorley-Lawson and Allday 2008). A critical feature of BL is the translocation of the *c-myc* proto-oncogene into one of the immunoglobulin loci (Figure 14), which leads to constitutively active *myc* expression and tumourigenesis. It is also frequently accompanied by p53 and other mutations (Crawford 2001). The translocation most likely occurs in germinal centres as the result of faulty somatic hypermutation and involves transfer of *c-myc* from chromosome 8 at band q24 to one of immunoglobulin loci at 14q32, 22q11 or 2p11 (Veronese et al. 1995, Crawford 2001) (Figure 15).



**Figure 14** – C-myc translocation visualised by FISH. Image courtesy of Monica Gostissa, PhD, PhD, PCMM/IDI, Children's Hospital Boston. Oncogenic *c-myc* translocation into the IgH locus is shown in green.

Other factors also increase the risk of the disease (Crawford 2001). In the regions of Africa and New Guinea where BL tumours are frequent and consistently EBV positive, BL is known as “endemic BL” (Figure 15) (Molyneux et al. 2012). In these BL-endemic and malaria-endemic regions, high prevalence of the malaria parasite is associated with higher incidence of EBV-positive BL, and antibody levels against EBV in African children increase the risk of developing BL later in life (Dethe et al. 1978, Piriou et al. 2012, Thorley-Lawson and Allday 2008). This suggests a causative role of EBV in areas of endemic BL (Molyneux et al. 2012).



**Figure 15** – Comparison of endemic BL distribution (upper panel) with global malaria parasite distribution (bottom panel). Sources: Dr Richard C. Hunt, “ONCOGENIC VIRUSES” at <http://pathmicro.med.sc.edu/>

However, in EBV-positive BL, viral expression is restricted and typical of latency mode I, normally associated with differentiated resting memory B-cells, therefore the possibility of

EBV as causal factor remains controversial (Thorley-Lawson and Allday 2008, Rezk and Weiss 2007). It has been proposed that in BL-endemic regions malaria increases BL risk through immunosuppression of T cells and by promoting viral reactivation resulting in higher EBV replication and load. As more B-cells are transformed and germinal centres are formed, the likelihood increases of one of them gaining a malignant mutation (Thorley-Lawson and Allday 2008, Crawford 2001). Malaria might thus increase the viral load, and also promote *c-myc* translocation by repetitive infection and antigen activation of B-cells (Thorley-Lawson and Allday 2008, Rasti et al. 2004). There is evidence that the viral contribution towards BL pathogenesis lies in the anti-apoptotic activity of the EBNA3 proteins that promote epigenetic silencing of proapoptotic Bcl2-like protein 11 expression in cells with active *c-myc* (Thorley-Lawson and Allday 2008). By this repressive mechanism, EBV-infected B-cells do not undergo apoptosis in response to EBNA-2 driven *c-myc* activation. However, this also means that they can survive with mutations such as the BL characteristic *c-myc* translocation (Thorley-Lawson and Allday 2008). Once they leave the lymph node, they stop the expression of most of latency genes, however unlike in normal EBV primary infection, the aberrant translocated *myc* enforces continuous proliferation (Thorley-Lawson and Allday 2008). Malaria could also promote translocation of *c-myc* independently, by inducing the enzyme cytidine-deaminase (AID) (Molyneux et al. 2012, Rasti et al. 2004).

A similar effect to malaria parasite manifesting in higher B-cell proliferation and viral load may be caused by HIV-induced immunosuppression (Ruf and Wagner et al. 2013, Hartman-Johnson et al. 2013) or, alternatively, HIV might contribute to oncogenesis directly (Grogg et al. 2007). However, only 40% of European HIV-related BL cases are EBV-positive and there is no conclusive evidence as to whether HIV infection increases the risk of BL additionally in endemic BL regions (Molyneux et al. 2012 Mutalima et al. 2008). This may implicate more than one underlying mechanisms. In HIV-related BL cases, which, unlike in endemic BL, can

be characterised by high levels of CD4 T-cells and lower EBV seropositivity, it has been proposed that chronic viraemia results in chronic mass activation of naïve B-cells and induction of AID (Molyneux et al. 2012). HIV-associated immunosuppression may then rescue cells with *c-myc* translocation in the absence of EBV. Recently, a study of Kenyan children born to HIV-infected mothers demonstrated they acquired EBV earlier and had higher viral loads, which may indicate that the virus can be a contributing factor (Slyker et al. 2013).

#### **1.4.2.9 Hodgkin's disease**

Hodgkin's lymphoma (HL) is one of the most common lymphomas in the world, with an incidence of 3 per 100,000 each year (Kuppers 2009). This type of lymphoma is specifically characterised by the low-level presence of abnormal B-lymphocytes called Hodgkin and Reed–Sternberg (HRS) cells, or lymphocytic and histiocytic (L&H) cells, depending on the morphological subvariety of HL (Kuppers 2009). These abnormal lymphocytes originate from B-cells losing the typical B-cell phenotype and expression and undergoing extensive reprogramming which alters global expression (Kuppers 2009). This is a unique feature of HL distinguishing it from other (non-Hodgkin) lymphomas. As a result HRS cells retain only receptors involved in helper T-cell co-stimulation, necessary for their survival in the germinal centre, and express markers of different haematopoietic cells including CD30 (tumour necrosis factor receptor superfamily member 8) and CD15 (Kuppers 2009). This might be essential in cellular interactions with non-tumour cells that form the majority of the tumour's tissue (Kuppers 2009). They have elevated expression of TFs that inhibit B-cell development such as Notch 1 and GATA (Kuppers 2009). Most importantly HRS cells have deregulated and active Jak–Stat and NF- $\kappa$ B pathways. The majority of genetic lesions typical of HL affect

both of these pathways, demonstrating their significance for HL tumourigenesis (Kuppers 2009). At the same time HRS cells are characterised by downregulated PAX5 – a TF responsible for B-cell lineage commitment (Kuppers 2009). The role of EBV in HL aetiology is unclear, however about 40% cases of classic HL are EBV-positive growing to 90% EBV-positive in paediatric HL and nearly 100% in HIV HL, with all cases invariably expressing the latency II programme (Kuppers 2009). It is thought that the viral proteins can substitute the effects of the pathogenic function of some mutant genes such as *TNFAIP3* and also contribute to B-cell survival in the germinal centre (Kuppers 2009). Interestingly, virtually all cases of HL in which HRS cells lack the receptors necessary for helper T-cell co-stimulation and survival are EBV-positive (Kuppers 2009). This suggests that the expression of viral proteins, LMP1 and LMP2, which substitute the survival signals in the germinal centre, might contribute to tumourigenesis. LMP1 is of particular interest as it constitutively activates the NF- $\kappa$ B pathway, whose essential role in HL has been highlighted by many authors (Crawford 2001, Hsu and Glaser 2000). A mechanism must exist, that prevents B-cells in the germinal centre from switching off LMP1 oncogene expression, even upon their exit into the circulatory system. This is possibly crucial to HL as constitutive abnormal activation of cellular pathways, mainly Jak-STAT and NF- $\kappa$ B is a central feature in contrast to other malignant lymphomas (Kuppers 2009). Nonetheless, it is still unclear what constitutes the critical step of HL outgrowth and whether reprogramming is involved or if it is a side-effect (Kuppers 2009).

### **1.4.3 Non B-cell tumours**

#### **1.4.3.1 Introduction**

For specific genetic backgrounds, EBV can infect other cell types. This has been occasionally seen in IM and frequently in CAEBV in which the virus can also infect T and natural killer (NK) cells (Rezk and Weiss 2007, Kimura et al. 2001). NK/T cell lymphomas, just as B-cell lymphomas, can range from a polymorphic to monoclonal malignant lines, with the latter arising from a pool of polyclonal EBV infected cells (Kimura et al. 2001, Young and Rickinson 2004). EBV infection has been found in cases of NPCs and gastric carcinomas, in which genetic predisposition possibly accompanied by early mutations might permit the virus to infect the epithelium (Young and Rickinson 2004).

#### **1.4.3.2 NK/T cell lymphoma**

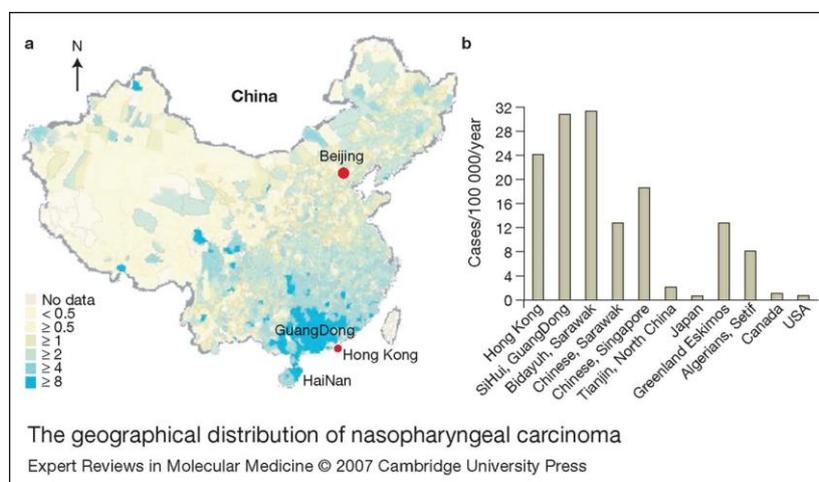
Extranodal NK/T cell lymphoma (or Angiocentric lymphoma of the nasal-type) is another type of tumour almost exclusively associated with EBV and predominant in East and South-East Asia, Central and Southern America, but rare in Western countries (Schmitt et al. 2011, Suzuki et al. 2008, Kuo et al. 2004, Motsch et al. 2013). It most often arises from a single clone of an NK cell, or less frequently (14% cases) T cells in extranodal sites within the nasal cavity, maxillary sinuses and palate, or in skin, testis, lungs and the gastrointestinal tract (Schmitt et al. 2011, George et al. 2012). The tumour has a tendency for dissemination and relapses, and is characterised by high mortality and a 5-year survival rate of below 50% (Schmitt et al. 2011, Takahashi et al. 2002). EBV genome copy number in serum or whole blood is indicative of treatment outcome, with high levels of virus associated with poor

disease outcome (George et al. 2012, Schmitt et al. 2011, Jaccard et al. 2009). EBV positive NK/T cell lymphomas normally express genes characteristic of type I latency or, less frequently, type II latency, similar to Reed-Sternberg cells in HL (George et al. 2012). Although the precise pathogenic mechanism is unknown it is thought that the virus contributes to the early stages of tumourigenesis (Fox et al. 2011, Huang et al. 2010, George et al. 2012). Oncogenic properties of LMP1 and its potential role in promoting survival of tumour cells might be of critical importance as blocking expression of LMP1 causes growth suppression of EBV positive NK cells (Murata et al. 2014). However, it should be noted that the expression of LMP1 in EBV positive NK/T cell lymphomas is not universal (George et al. 2012). Most importantly it is still unknown how the virus manages to infect mature T and NK cells, which lack the crucial CD21 co-receptor (George et al. 2012). Such cells are observed in tonsils of patients undergoing primary EBV infection (George et al. 2012). The fact that high prevalence is restricted to certain populations indicates that genetic and local environmental factors contribute to tumourigenesis (George et al. 2012). EBV-positive T and NK cell malignancies also include other, rare disorders like angioimmunoblastic lymphadenopathy (a peripheral T-cell lymphoma) and NK/T-cell granular lymphoproliferative disorder (Cohen 2000, Hiraki et al. 2001, Hsu and Glaser 2000, Delecluse et al. 2007), which has recently re-classified as a separate entity in lymphoma classification (Gattazzo et al. 2013).

#### **1.4.3.3 Nasopharyngeal carcinoma (NPC)**

The non-keratinizing undifferentiated subtype of NPC (WHO type III) represents 85% of NPC in high risk regions and is associated in virtually all cases with EBV. In China and other NPC endemic regions the EBV-association is high for both non-keratinizing NPC subtypes,

WHO types II and III, as well as keratinizing NPC (Yoshizaki et al. 2012, Shair et al. 2009, Shah and Young 2009, Young and Rickinson 2004, Wildeman et al. 2012). The virus, expressing latency II genes, has been proposed to play a critical role in cancer development, however additional genetic and environmental factors are also thought to contribute (Jia and Qin 2012, Hildesheim and Wang 2012, Louis et al. 2010). This is because NPC follows an unusually restricted and highly localised distribution, being generally rare world-wide (with an incidence of 1 case per 100,000) (Marquitz and Raab-Traub 2012), however much more frequent in South-East Asia with an incidence of 6.5 cases per 100,000. This rises to 20-31 per 100,000 in certain Chinese cities and other places in South-East Asia, particularly among the Cantonese ethnic group (Jia and Qin 2012) (Figure 16).



**Figure 16** - NPC distribution in China. (reproduced from Tao and Chan 2007)

Family history of cancer, especially NPC, increases the risk 4 to 10 times (Jia and Qin 2012). Studies of risk in migrant populations indicate both regional environmental factors and genetic predispositions are important in NPC development (Jia and Qin 2012). In the past decade variants in HLA Class I genes, and several other candidates: *RAD51L1*, *MDM2*, *MMP2* and *TP53*, were identified as associated with NPC, however the modest size of the discovery cohorts and no replication studies undermine the significance of the findings

(Hildesheim and Wang 2012). The HLA locus has been the most consistent finding across all studies and the largest GWAS study also identified other potential associations including *TNFRSF19* (Hildesheim and Wang 2012). A model has been proposed in which environmental factors induce EBV reactivation and DNA damage, and act together with genetic predispositions to enable EBV to infect epithelial cells where it cannot differentiate, and where constitutive expression of LMP1 induces carcinogenesis directly (Thorley-Lawson 2001, Jia and Qin 2012, Dawson et al. 2012). Thus, in contrast to B-cell lymphomas, EBV is thought to be the direct cause of the disease (Shah 2009, Young and Rickinson 2004). A treatment with EBV-specific cytotoxic T cells has been proposed, which would target viral proteins involved in tumourigenesis including EBNA1, LMP 1 and LMP2 (Louis et al. 2010).

#### **1.4.3.4 Gastric carcinoma**

5 to 10% of adenocarcinomas and 90% of lymphoepithelial carcinomas of the stomach are characterised by EBV infection of the epithelium in which latency type I or type II genes are expressed. LMP1 expression is present inconsistently and at low levels or, in the case of adenocarcinoma, absent altogether (Shah and Young 2004, Delecluse et al. 2007). LMP2 expression can also be absent (Fukayama et al. 1994). The absence of LMP1 and EBNA2 expression (both known to be the key oncogenes) in latency type I tumours has led to a hypothesis that other factors, most likely somatic mutations, are the direct cause of cancer (Fukuyama et al. 1994). Virtually all cases of GC are characterised by monoclonal EBV genomes and thus the virus is thought to epigenetically drive the alteration of global expression including inactivation of tumour suppressors (Fukuyama et al. 1994).

#### **1.4.4 Other EBV-associated disorders**

EBV DNA and latency proteins may also play a role in pathogenesis of smooth-muscle tumours in transplant recipients (Hiraki et al. 2001, Cohen 2000, Delecluse et al. 2007). The virus is also a direct cause of oral hairy leukoplakia in patients with acquired immunodeficiency syndrome (AIDS) manifesting in benign lesions of the tongue caused by lytic viral infection of the differentiated squamous epithelial cells (Crawford 2001, Delecluse et al. 2007).

#### **1.4.5 EBV-related autoimmune diseases**

EBV has also been implicated in autoimmune diseases, notably Multiple Sclerosis (MS), Systemic Lupus Erythomatosus (SLE) and Rheumatoid Arthritis (RA). The body of evidence indicating an association between EBV and MS includes almost universal seropositivity for the virus in MS patients, similar latitudinal distributions of IM and MS and strong evidence that people with MS are more likely to have undergone IM in the past (Ramagopalan et al. 2010, Lossius et al. 2012). Some studies suggested increased MS risk in association with higher levels of antibodies against the EBNA<sub>s</sub> (Ramagopalan et al. 2010). Higher seropositivity and anti-EBV antibodies are also associated with increased risk of SLE (Lossius et al. 2012). Additionally, the epitope-specificity of the anti-EBV antibodies differs between healthy individuals and SLE patients (Lossius et al. 2012). Elevated anti-EBNA titres are also a feature of RA, but do not predate the disease (Lossius et al. 2012). Some studies have suggested that EBV DNA load in peripheral blood mononuclear cells (PBMC) may be higher in patients with a clinically isolated syndrome (CIS) or relapsing–remitting MS (Ramagopalan et al. 2010). Higher EBV load has also been demonstrated in RA and SLE patients (Lossius et al. 2012).

Several theories have been proposed to explain these associations (Pender 2011, Ascherio and Munger 2007). For example, one asserts the autoimmune response is caused by cross-reactivity of some T-cells to both EBV antigens and the host's cellular peptides, for instance myelin-derived (Pender 2011, Yasui et al. 2008, van Noort et al. 2000). According to a variant of this hypothesis, low expression of a heat shock protein called  $\alpha$ B-crystallin which is normally absent in most human tissues leads to its low immune tolerance and can result in autoimmune reaction against  $\alpha$ B-crystallin derived from oligodendrocytes and demyelination after CNS inflammation (Lucas et al. 2011).

Another suggestion is that EBV might promote survival of auto-reactive B-cells that in immune-compromised patients with appropriate genetic predispositions, provide survival signals for auto-reactive T cells driving autoimmune reaction (Pender 2011). Subsequent epitope spreading, in which the immune system's becomes sensitized to myelin and other CNS antigens released as a result of tissue damage during the initial auto-reactive response, has also been proposed as a possible explanation for certain features of MS and experimental autoimmune encephalomyelitis (autoimmune disease induced experimentally in animal models by exposure to myelin antigens) (Pender 2011; Vanderlugt and Miller 2002, Pender and Greer 2007, Pender and Wolfe 2002,). Genetic variation of EBV receptor genes (Simon et al. 2007) or EBV strains (Brennan et al. 2010) and co-infection with EBV and a retrovirus in genetically susceptible individuals (Munch et al. 1998) have also been speculated to cause or increase the risk of MS but no specific mechanism was proposed and no evidence provided.

Another mechanism has been proposed whereby EBV-infected B-cells present in active white matter MS lesions secrete EBER transcripts which bind toll-like receptor 3 (TLR3) displayed on the surface of dendritic cells. This elicits interferon-alpha secretion which stimulates and propagates immune responses exacerbating the inflammation (Luenemann 2012, Tzartos et

al. 2012, Meier et al. 2012). This theory does not explain the direct cause of MS however it suggests how the virus might activate and maintain neuroinflammation in MS lesions that are already active (Tzartos et al.2012).

A strong epidemiological correlation between IM and MS was the reason why IM has been proposed as a causal factor for MS (Thacker et al. 2006, Lucas et al. 2011). Because the number of CD8+ T cells normally declines as age increases, the primary CD8+ T cell deficiency will increase as each patient ages. This could explain the age-dependent accumulation of disability seen in MS patients. Prolonged high EBV load could lead to T cell exhaustion, aggravating the CD8+ T cell deficiency further (Pender 2011). Alternatively, inflammation and polyspecific B-cell activation accompanying the autoimmune response could themselves be responsible for features like higher EBV load in patients. This however does not explain high antibody titres prior to the occurrence of an autoimmune disease or the link between IM and MS.

There is evidence that it is the IM rather than the EBV infection alone, which is associated with malignancies like HL and MS (Hadinoto et al. 2008). Past history of IM increases the risk of MS twofold (Ramagopalan et al. 2010). Since up to 25% of memory B cells can be wiped out within a week, this means an extensive commitment of the immune system's memory to cellular and viral antigens (Hadinoto et al. 2008). This action could be what is responsible for the deregulation of the immune system and the predisposing factor for the IM-associated diseases. EBV would thus be a co-factor rather than the primary cause of the disease (Hadinoto et al. 2008).

## **1.5 Host genetics and EBV Infection**

There is inherent heterogeneity in EBV infections manifesting most importantly in different levels of viral load in B-cells and the total number of infected B-cells (Hiraki et al. 2001, Crawford 2001, Straus et al. 1993, Dimmock et al. 2007, Hsu and Glaser 2000, Zandman-Goddard et al. 2009). LCLs can vary substantially in the number of viral copies per cell (from 1-2 up to 800 and possibly more) and genetic determinants have been suggested as the underlying cause (Caliskan et al. 2011, Sugden et al. 1979). However, factors like EBV copy number can in turn significantly affect host's expression (Choy et al. 2008). The response of the host to EBV infection can vary substantially from an asymptomatic subclinical infection to persistent chronic infectious mononucleosis or lymphoproliferative disease (Crawford 2001, Straus et al. 1993, Rubicz et al. 2013). This could be to some extent determined by environmental factors such as viral dose, presence of the malaria parasite or temporary immuno-suppression, however the genome of the host also plays a vital role in modulating the susceptibility and the course of infection (Rubicz et al. 2013). Genetic determinants are responsible for predispositions to certain EBV-related diseases (like NPC, endemic to South-East Asia populations) or EBV-immunity, as seen in individuals Bruton's agammaglobulinemia, where B cells do not express the CD21 receptor (Goldman 1999). Various risk loci have been identified for HL, NPC, IM and MS (Rubicz et al. 2013). Some of them also affect EBV biology. For instance, genetic variants of the human IL-10 gene have been reported to modulate virus-host interactions as they are positively correlated with earlier onset of primary EBV infection in children (Calderwood et al. 2007). Other polymorphisms proximal to IL-10 have at the same time been associated with higher SLE risk in African Americans; also certain HLA class I alleles are significantly associated with IM rather than an asymptomatic infection and at the same time HLA I markers (D6S265 and D6S510) were associated with EBV-positive lymphomas in patients versus controls (Williams et al. 2004,

Cameron et al. 2006, Hochberg et al. 2004, Hadinoto et al. 2008). There is also some evidence that heterogeneity in response to IM is determined by the host's ability to mount an immune response, but could also be associated with the viral dose (Niens et al. 2007, Skibola et al. 2007). The T-cell depletion hypothesis links higher viral dose with a more severe disease phenotype (Pender 2011), consequently it is likely that EBV copy number and the level of expression of key transforming and oncogenic viral transcripts affect cellular phenotype and disease course. These could themselves be pre-determined by genetic variants that independently appear as disease risk loci for EBV diseases. Some studies of viral expression in animal models support this hypothesis (Kieff et al. 2010). In this context, a global survey of viral copy number and latency eQTLs integrated with GWAS risk loci determined for EBV-related conditions could prove very informative. Recently, a study attempting to carry out such an analysis identified an EBV-specific anti-EBNA1 antibody level QTL in the HLA class II locus (in *HLA-DRB1* and *HLA-DQB1*) in a panel of over 1300 Mexican American family members (Rubicz et al. 2013). As previously discussed, higher EBNA-1 antibody levels are a feature of MS, SLE, in addition to NPC and BL (Rubicz et al. 2013). Interestingly, *HLA-DRB1* expression level correlated with anti-EBNA1 expression level and these QTLs overlapped with risk loci for NPC and HL. One of the top EBNA1 QTL SNPs turned out also to be a top significantly associated locus in a Spanish SLE cohort, while another had been previously associated with MS (Rubicz et al. 2013). Additional human eQTLs were linked to the anti-EBNA1 QTL SNPs. Authors of the study speculated that EBNA-1 protein levels could be seen as a proxy for EBV viral load, which is regulated by the associated genetic determinants responsible for higher risk of autoimmune diseases and tumours (Rubicz et al. 2013).

There has been much speculation in the literature over the possibility of different EBV strains (not just the two main types distinguished by EBNA2 polymorphisms, but more specific

substrains with other polymorphisms in EBNA1, EBNA2, and LMP1 and some certain lytic genes, sometimes more restricted geographically) potentially exploiting the genetic susceptibilities inherent in their host populations and promoting EBV-associated diseases, for instance chronic lymphocytic leukaemia (polymorphisms in matrix metalloproteinase 9) or NPC (Hjalgrim et al. 2010, Da Silva et al. 2007, Ikegaya et al. 2008, Chang et al. 2009). However only a few underpowered studies have addressed this issue with inconclusive evidence so far. Studying viral transcripts and proteins offers the key to a better understanding of their functions in vivo, and thus their relevance in viral biology and disease. So far, evidence suggests that examining the association between the host genetic variation and differential viral load or viral protein expression level could yield new insights into the mechanisms of EBV infection by suggesting the molecular links between cellular and viral pathways. It is now clear that latency genes play an important aetiological role in EBV-related diseases and different disease outcomes are likely to be mediated by different levels of EBV expression and viral load. Consequently, it would be beneficial to fully describe the genetic regulatory effects that influence the host response to infection and determine viral uptake and activity.

## **1.6 Genetic determinants of gene expression**

### **1.6.1 Transcript level as mediator of genetic effects on the phenotype**

There is significant variation in transcript expression between individuals and populations across cell and tissue type (Cookson et al. 2009, Gilad et al. 2008). This includes LCLs where analysis of cell lines established from family trios and monozygotic twins demonstrated

significant heritability (Cheung et al. 2010). Stranger et al. (2012) investigated global expression in 30 Caucasian and 30 Yoruban trios as well as 45 unrelated Chinese and 45 unrelated Japanese and found that 10% of expressed transcripts in Caucasians and 13% in Yorubans to have heritability of over 0.2, and between 17% to 29% transcripts had significant expression differences between any two populations (Stranger et al. 2012). Coding DNA sequence variants are transcribed to produce functional RNA such as mRNA which is translated into proteins, rRNA which constitutes part of ribosomes, or microRNAs which regulate mRNA degradation (Gilad et al. 2008, Freedman et al. 2011). Coding variants alter the molecular structure of the gene product directly which may influence transcription efficiency, transcript stability and, if the variant is non-synonymous, also affect protein structure and functionality (Freedman et al. 2011). This can in turn affect the expression of other genes if the altered protein acts as a co-repressor or activator. For instance mutations to p53, a master TF, caused variation in binding efficiency and transactivation of its target response elements (Resnick and Inga 2003). Non-coding variants affect DNA elements controlling gene transcription like promoters, enhancers and other TF binding sites and are mostly responsible for differences in gene expression levels. Although not altering the molecular structure directly, by altering the level of transcript and the amount of protein produced non-coding variants can still affect the phenotype, for instance through protein level threshold dependent effects (Gilad et al. 2008, Freedman et al. 2011). Some of the most striking examples are provided by the numerous haplo-insufficiency syndromes like Holt-Oram syndrome, lymphedema-distichiasis syndrome or Waardenburg syndrome, in which deleterious mutations to TFs reduce their expression by approximately 50% and result in disorder often with pleiotropic effects on the phenotype (Seidmann and Seidmann 2002).

### 1.6.2 Mapping gene expression as a quantitative trait

Searching for association between transcript abundance and genetic variation offers a means to discover the regulatory effects that the genome exerts on the transcriptome. Studies of the genetics of transcription adopt the level of transcript expression as the quantitative phenotype of interest. Genetic variants that are found to influence this phenotype are hence termed expression quantitative trait loci (eQTLs). (Damerval et al. 1994) The regulatory effects of eQTLs can be broadly grouped into two categories depending on the distance of the variant from the gene whose transcription it controls (Gilad et al. 2008). An alternative classification is based on the eQTL's ability to affect the expression of a single allele versus both parental alleles (Gilad et al. 2008). Cis eQTLs, which account for most variability in expression, are typically located within 100-500kb of the transcription start site (TSS), although there is no single consensus on what constitutes a local versus distant interaction (Cookson et al. 2009, Freedman et al. 2011). The majority of cis eQTLs are located less than 100kb from the target gene's transcribed sequence and cluster symmetrically around its TSS (with 33% of *cis* eQTLs located within 10kb from TSS), while the more distant regulatory loci are fewer and have a smaller effect (Cookson et al. 2009, Veyrieras et al. 2008). A similar cluster of strong regulatory SNPs is also present immediately before the transcription end site (TES) and decreases rapidly beyond it (Veyrieras et al. 2008). Cis eQTLs often reflect the effects of promoters and proximal enhancers and insulators. eQTLs located further from the TSS, sometimes on different chromosomes, are termed trans-eQTLs. These may be caused by polymorphisms affecting transcription factor genes and are thought to be more frequent yet less effective than cis-eQTLs (Cookson et al. 2009, Freedman et al. 2011). However they can also be located in enhancers and other regulatory sequences who can act at distance with help of CTCF and or DNA-looping and structure-altering proteins (Watson et al. 2008, Phillips and Corces 2009).

Transcript levels are mostly determined by the efficiency of transcription, which itself is regulated by specific combinations of TFs and cofactors and their affinity to bind specific motifs within regulatory sequences like proximal and distal promoters and enhancers (Veyrieras et al. 2008). Post-transcriptional modifications such as splicing and transcript stability play a secondary role. However most eQTLs within an ORF are located in exonic rather than intronic sequences, which indicates that some of them might also be influencing the stability and degradation rate of mRNA (Veyrieras et al. 2008). The epigenetic modifications of histones and CpG methylation also influence transcription through altering the quaternary structure and conformation of the DNA and are particularly important in control of tissue-specific expression (Veyrieras et al. 2008, Cookson et al. 2009, Wang 2013). Most eQTLs are found in regions of open active chromatin, accessible for TFs and detectable with DNase-seq and Formaldehyde Assisted Isolation of Regulatory Elements technique (FAIRE) (Wang 2013). Thus eQTLs, TF-binding sites and regions of activate chromatin are often co-localised (Wang 2013). Cell type-specific regulatory sequences, subjected to epigenetic silencing, are usually located away from the TSS, where the DNA modifications may vary more between the cell types (Wang 2013).

### **1.6.3 Genetic mapping and GWAS**

Initially eQTL (and other phenotypic) studies were based on linkage and co-transmission of trait-associated chromosomal segments from parents to offspring and co-segregation of the trait and transmitted fragment in pedigrees, and conducted without the use of genome-wide genotypic data (Gilad et al. 2008, Mackay et al. 2009). This method relies on the principle of meiotic recombination and independent assortment, whereby closely located genes should be more frequently co-transmitted and result in correlation between a transmitted chromosome

segment and a phenotype (Gilad et al. 2008). Linkage studies, while powerful to detect solitary strong-effect risk loci causing monogenic Mendelian disorders, lack the sensitivity to detect small effects of co-transmitted risk loci since they cannot separate them from the “background noise” effects caused by other (shared) risk loci and environmental effects without greatly increasing the required numbers of studied families, particularly if the small-effect risk locus is common and its transmission difficult to trace in pedigrees (Kruglyak 1997, Hirschhorn and Daly 2005). Linkage studies have also serious limitations in mapping resolution (because of large LD-blocks shared by related individuals) and as a result have been largely replaced by a population-based association approach that studies associations in unrelated individuals (Gilad et al. 2008). This approach enables to investigate common risk loci and is more powerful given the same sample size (Stewart and Cerise 2013). It can be subdivided into a binary case-control and quantitative design (Bush and Moore 2012). The underlying argument is that a genotype/allele dose corresponds to trait quantity or that the genotype/allele frequency difference between cases and controls corresponds to the disease incidence, which should provide evidence for statistical association, whereas the quantitative approach (Cardon and Bell 2001, Bush and Moore 2012). Such studies collect large unrelated samples from a population and have greater power to detect small-effect common variants at higher resolution, particularly when combined with whole-genome association testing. Whole genome expression studies can be coupled with genome-wide assays of genetic variation to give an overall global picture of regulatory effects, improving the understanding of gene and protein function, and pathways (Gilad et al. 2008, Mackay et al. 2009). By replacing transcript expression as the phenotype of interest with disease status or some other disease-related characteristic, it is possible to gain insight into genetic determinants influencing disease susceptibility and clinical outcome. Such studies, termed genome-wide association studies (GWAS) have been at the forefront of medical genetics in

the past decade and identified novel disease risk loci responsible for a varying fraction of the disease heritability. Such risk loci provide candidates for further functional investigation and might help to understand the causes of the disease. Research on Inflammatory Bowel Disease provides a good example of progression from multiple GWAS risk loci (numbering over 100 in 2011) to underlying causal mechanism (Dubinsky et al. 2011, Khor et al. 2011). This condition, which is usually subdivided into chronic Crohn's disease and ulcerative colitis, is partly heritable and its risk loci overlap not only between the two types of the disease mentioned above, but also for certain other diseases like RA or susceptibility to *Mycobacterium leprae* infection. Pathway analysis identified that many of genes harbouring risk loci share a common functional background converging on cellular pathways responsible for epithelial barrier and transport function, epithelial restitution, microbial defence, and several others (Khor et al. 2011). Functional studies conducted using cell lines and mouse models indicated that in several cases investigated mutant alleles lead or may lead to, among others, abnormal barrier permeability enabling microbial incursion, an imbalance between pro-inflammatory and anti-inflammatory cytokines and abrogation of microbial homeostasis in the intestine resulting in inflammation (Khor et al. 2011). Identification of potential molecular targets in turn lead to the proposal of new anti-cytokine therapies for the treatment of the disease, for instance with anti-TNF antibodies or *Lactobacilli* that produce TNF-specific [nanobodies](#), which are currently under investigation (Kaser et al. 2010, Neurath 2014). Global eQTL scans, when integrated with data provided by GWAS studies, provide more information on the molecular basis and biology of disease susceptibility linking risk loci to genes and pathways (Cookson et al. 2009, Freedman et al. 2011). This is particularly relevant for complex diseases for which multiple small-effect loci causing differences in gene expression could act as the predisposing factors, rather than single, highly significant SNPs directly altering protein structure and function (Choy et al. 2008). Indeed for most complex

diseases the associated variants fall largely into non-coding sequences and consequently a significant proportion of them should also be expected to influence gene expression (Freedman et al. 2011). It is apparent that for multifactorial diseases many risk loci are likely rare variants, which are challenging to detect without large fully sequenced cohorts. To date 1738 GWAS publications have discovered 11,751 SNPs associated with a specific disease or condition, however most of the findings lack the subsequent functional investigation necessary to determine causality and nature (GWAS catalog at [www.genome.gov](http://www.genome.gov) , accessed on 27.11.2013).

#### **1.6.4 eQTLs in post-GWAS analysis**

The ultimate aim of establishing causality poses a methodological obstacle as there are no clear criteria for variants located in non-coding sequences, however one of the proposed initial steps in functional annotation of such disease risk loci to consider eQTLs to identify their potential regulatory effects (Freedman et al. 2011). Distribution and overlap of potential eQTLs and risk loci should then help to prioritise the latter for further functional investigation. In this way a candidate disease risk locus can be quickly linked to a molecular phenotype (Gilad et al. 2008). Other steps could include re-sequencing and fine-mapping of associated SNPs and analysis of surrounding linkage disequilibrium (LD) blocks. Combined studies of methylation, histone modifications and TF-abundance can also help to narrow down likely functional variants (Freedman et al. 2011). There is a body of evidence suggesting that methylation levels of CpG islands are to an extent genetically determined and such methylation QTLs are associated with histone/chromatin modifications and gene expression (may overlap with eQTLs), and also with changes in disease risk (Banovich et al. 2014). A simple follow-up assay could correlate allele genotypes at risk loci with expression

levels of corresponding candidate transcripts. In heterozygous individuals allele-specific expression assays could also be conducted looking at potential difference in expression between the two copies of the gene (Gilad et al. 2008, Freedman et al. 2011). The study could then be expanded to include functional assays for TF-binding, chromatin immunoprecipitation (ChIP), electrophoretic mobility shift assay (EMSA) and animal knock-out models (Freedman et al. 2011). Since genes co-operate and interact at the protein level and transcript abundance is sometimes a polygenic trait itself, expression from candidate eQTL loci could be integrated with global expression and protein interaction data in pathway analysis to identify biologically related networks of genes participating in common cellular pathways important for the disease of interest (Gilad et al. 2008, Freedman et al. 2011). This type of analysis has been done by Cookson et al. (2009) who conducted a case-control asthma GWAS in a total sample of over 2000 individuals, and investigated global gene expression in 400 asthmatic children and their healthy siblings. They found a significant asthma risk locus, whose genotype, after global expression profiling, correlated with expression levels of a nearby gene encoding a transmembrane endoplasmic reticulum protein, Orosomucoid like 3 (ORMDL3). The finding has been since replicated in several other cohorts (Knight 2009), and ORMDL3 became subsequent subject of functional studies, some of which proposed mechanistic models explaining the disease association (Hsu and Turvey 2013, Miller et al. 2014, Carreras-Sureda et al. 2013, Cantero-Recasens et al. 2010)

### **1.6.5 Studies of human and model organism eQTLs**

The ability to detect eQTLs is dependent on the strength of their effect and samples size, and the extent they tend to cluster together (Mackay et al. 2009). Overall, eQTL studies in both humans and animal models indicate that *trans* eQTLs with strong effects on multiple

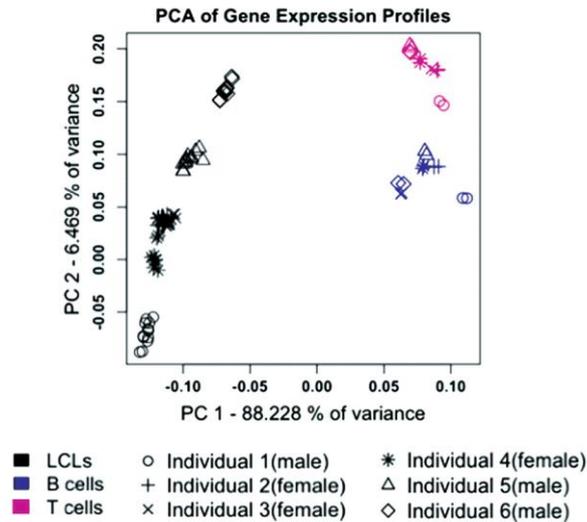
transcripts are much less frequent in humans than in other model organisms (Gilad et al. 2008, Mackay et al. 2009). Studies of model organisms, such as *Arabidopsis* or yeast, indicate that often many loci with small effects can cluster together. If the eQTLs have opposite effects, this can make them difficult to map, especially with low-resolution genotyping methods (Mackay et al. 2009). Typically, transcript expression is correlated and multiple transcripts will form co-expression networks dependent on common factors controlling their levels, which reflect their related function and participation in similar cellular functions (Mackay et al. 2009).

### **1.7. LCLs as a model of EBV-infection**

The transforming properties of EBV have provided a useful tool to study human genetics. EBV undergoes very efficient lytic replication in cotton-top tamarine's B-lymphocytes (causing lethal lymphoma *in vivo*). Human B-cells inoculated with virion-containing tamarine cell culture (B95-8) supernatant convert to immortalised LCLs and grow stably in suspension with a relatively low mutation rate of 0.3% offering a reliable means of DNA-storage readily available for genetic screening (Diehl et al. 1978). This advantage has made LCLs a popular platform in a wide range of molecular experiments, including many expression studies. As a resource they are cheaper than many primary tissues and are readily available from a number of cell repositories and biobanks. Most notably, Coriell Institute's Cell Repositories provided LCLs sourced from multiple worldwide populations used in large scale analysis of genomic variation by the HapMap consortium, HapMap 3 project, and 1000 Genomes Project. The International Hap Map project was launched as an international effort to provide genome-wide maps of at least one million SNP resolution from different human populations, which could be used in GWAS (TIHMP 2003). The project genotyped 1,6

million common SNPs at 5% MAF in 1,184 individuals from 11 populations (Altschuler 2010). The 1000 Genomes project aims to sequence more than 1000 individuals from several distinct populations to survey human genetic variation world-wide, including rare variants with minor allele frequency (MAF) of down to 0.1%. LCLs used in HapMap and 1000 Genomes projects have subsequently been used in expression, RNA-seq and GWAS studies as well as functional studies conducted under the auspices of the ENCyclopedia Of DNA Elements (ENCODE) Project (Dunham et al. 2012). ENCODE is an international consortium of scientists working to identify all functional elements in the human genome (including genes, promoters, enhancers, repressors/silencers, exons, origins of replication, sites of replication termination, transcription factor binding sites, methylation sites, deoxyribonuclease I (DNase I) hypersensitive sites, chromatin modifications and multispecies conserved sequences) using high throughput molecular techniques (Dunham et al. 2012, The ENCODE Project Consortium).

Studies in LCLs are nonetheless prone to a number of important limitations and confounders. Primarily, B-lymphocytes do not faithfully reflect the biology of other tissue types and therefore might not be an adequate model for tissue-specific expression in many complex diseases of non-lymphatic origin. Moreover, because they are derived from B-cells, and due to EBV's pleiotropic regulatory effects, the expression profile of LCLs does not always ideally reflect that of other lymphocytes, see Figure 17 (reproduced from Caliskan et al. 2011)



**Figure 17** – B-cell, T-cell and LCL expression PCA plot.

This is why recent projects, such as Genotype-Tissue Expression run by the US National Institutes of Health Common Fund, focus on multiple primary cell lines to take into account tissue-specific expression (Lonsdale et al. 2013). Additionally, after a certain period of culture growth, LCLs reach a stage at which they become highly prone to aneuploidies and somatic mutations. Therefore primary stocks must routinely be subjected to cryopreservation after each thaw or passage; the cells become discordant upon reaching 40-50 or more passages (Mohr 2010). Other non-genetic factors and confounders like baseline growth, EBV titre used for inoculation, EBV copy number and individual response to EBV infection are also likely to affect the LCLs and should be taken into account (Choy et al. 2008).

Paradoxically some of these limitations become an advantage when LCLs are used specifically to model naïve B-cells activated by EBV. Although their global expression profile may differ from many other tissues due to EBV infection, for the very same reason it should faithfully reflect the earliest stages of infection and viral latency III in naïve B-cells.

## 1.8. Previous work and rationale for the thesis

### 1.8.1 LMP1 effects on human expression in hypoxia study

As previously discussed, EBV interacts with host genome and proteome altering cellular pathways and gene expression in multiple ways. A study of global hypoxia response eQTLs in a panel of Yoruban HapMap LCLs measured the levels of the viral LMP1 oncogene in order to regress out its effect as a potential confounder (Mohr 2010). This is because LMP1 affects gene expression globally and has been reported to up-regulate the levels of hypoxia inducible factor 1-alpha (HIF-1a) in lymphoblastoid as well as nasopharyngeal epithelial (EBV negative) cell lines *in vitro*, consequently activating HIF-responsive genes (Wakisaka et al. 2004, Mohr 2010). HIF is a master TF that regulates glycolytic enzymes responsible for ATP production under low oxygen conditions and a global regulator of monocyte and neutrophil immune response in hypoxic inflammatory environment (Zinkernagel et al. 2007). It has been proposed that LMP1 is associated with increased HIF-1 expression as it prevents its degradation (Zinkernagel et al. 2007 Benders et al. 2009 Kitagawa et al. 2013). This *in vitro* effect was however not replicated in NPC tissue and has been recently questioned (Bedners 2009, Chiang et al. 2013), but at the same, a similar stabilising effect of EBNAs has been reported (Morinet et al. 2013). If the LMP1 effect is real, it could potentially become a serious confounding factor altering global expression and therefore it had to be adopted as a covariate in the hypoxia study.

The study showed that LMP1 levels affect transcription in 58 HapMap LCLs from the Yoruban panel (Mohr 2010). Covariate regression (using an additive linear model with 5 known covariates including LMP1 transcript level), indicated that 6% of probes under hypoxia, and 7% under normoxia (Illumina HumanWG-6 Expression BeadChips (v3.0) )

were associated with LMP1 levels (Mohr 2010). These probes corresponded to 154 genes (1.7% total). However, it also became evident that the level of the viral oncoprotein, crucial to the course of viral infection, could itself be affected by host genetic determinants. An association close to the nominal genome-wide significance of  $5 \times 10^{-8}$ , was found under normoxic conditions between the LMP1 transcript and a specific host SNP (rs1913243, chromosome 3) located in an intergenic region between *CMYAI* (cardiomyopathy associated 1) and *CX3CRI* (chemokine C-X3-X motif receptor 1) (Mohr 2010). The SNP is in linkage disequilibrium (LD) with other neighbouring SNPs, some of which could be located within cis regulatory elements and affect transcription locally. Expression levels of proximal genes were not investigated, however the reported SNP was shown to be associated with mRNA abundance of *PPPDEI* (chromosome 1) – an indication of a possible *trans* effect. Additionally, a GWA test for the difference in LMP1 levels between hypoxia and normoxia, yielded a significant (P-value of  $8.6 \times 10^{-10}$ ) candidate on chromosome 19 (rs2866464), proximal to *C19orf2* – a gene reported to be involved in response to hepatitis B virus X protein (Mohr 2010). In summary, the Mohr study confirmed some of the theories on human-EBV interactions found in the literature. It clearly suggested that genetic determinants can affect the abundance of viral latency genes and identified putative regulatory variants for further investigation.

### **1.8.2 Rationale**

EBV biology is complex and tightly intertwined with human B-cell development and multiple cellular signalling pathways. There is consistent evidence that viral copy number and viral transcript levels can exert a quantitative effect on gene expression in infected B-cells (Choy et al. 2008, Mohr 2010, Caliskan et al. 2011). On the other hand, human genetic

variation may affect the course of EBV latent infection through moderation of the uptake and survival of the virus. This is reflected in the variable percentage of infected B-cells in different individuals (Thorely-Lawson 2001) and by differing viral copy number or the level of viral transcripts and proteins (Caliskan et al. 2011, Rubicz et al. 2013, Mohr 2010, Choy et al. 2008). The numerous interactions of viral proteins with human regulatory TFs that exert pleiotropic effects on gene expression, point to the possibility that transcription of viral genes and EBV persistence in B-cells (as represented by the viral copy number per cell) is subjected to the regulatory influence of human genetic variation. Genetic variants may affect the expression, transcript stability, or functionality and structure of the cellular machinery that the virus utilises to its own advantage. There is direct evidence to support this claim, and previous work conducted at the Wellcome Trust Centre for Human Genetics provided preliminary data that justified a larger study by revealing associations between expression levels of a key viral oncogenic transcript and human genetic variation (Section 1.8.1).

To date, studies in this field have investigated only a single correlation or a single viral quantitative trait at a time. Consequently, a research study that would further address the issue of genetic determinants of EBV infection and explore the findings of the Mohr study would be a valuable contribution to the literature. This is especially true in the broader context of EBV latency which would encompass the latency transcripts as well as global host expression. Due to the numerous contributions of EBV latency proteins to human disease, such work could prove informative, reveal novel regulatory mechanisms, shed more light on viral-host interactions and potentially indicate risk loci that increase disease susceptibility.

Therefore the aim of this project is to investigate the consequences of human genetic diversity for the activity of EBV in LCLs.

### 1.8.3 Hypothesis

The underlying assumption of this thesis is that SNP differences can, through multiple regulatory mechanisms, affect viral gene expression and copy number retention just as they affect the expression of human genes and contribute to inter-individual and inter-population phenotypic differences. Such effects are measurable through mRNA and DNA quantification methods and the current study will aim to validate this hypothesis using a cohort of LCLs from several hundred individuals and investigating all the EBV latency transcripts and copy number in search of potential regulatory loci which could then be subject to further research.

The current study will quantify the viral phenotypes of interest across panels of human LCLs and identify and map their genetic determinants through association tests. Candidate associations obtained from this analysis will be subsequently prioritised by the strength of association, function and relevance to EBV biology. Then, the results will be integrated with human expression data to search for SNPs that show regulatory effects on expression of both EBV and human transcripts and as such, might indicate viral-host interactions and their function. The top eQTLs for both human and viral expression, will be subjected to a replication follow-up in naïve and stimulated B-cells as well as newly transformed LCLs derived from the same individuals.

EBV has been linked to multiple human disorders due to oncogenic, growth-promoting properties of its latency proteins which hijack innate pathways responsible for B-cell maturation and modify them to ensure long term viral persistence and replication (Crawford 2001, Hiraki et al. 2001). It therefore follows that any genetic variant influencing transcription and transcript levels of EBV latency proteins, particularly its key oncoproteins, might potentially be a risk factor in EBV-related diseases and thus provides a subject for further investigation.

## **1.9. Project's aims**

### **Aim 1: To validate the association of LMP1 with rs1913243**

The first aim of the project is to validate the LMP1 eQTL (rs1913243) from the Mohr study by investigating possible regulatory effect that the candidate eQTL, or any of the neighbouring SNPs in high LD, could exert on the transcription levels of neighbouring genes. This will encompass assaying the same panel of Yoruban HapMap LCLs for expression of chosen transcripts by qPCR, followed by regional association testing. Candidate transcripts will be selected on their proximity and potential to interact with proteins of EBV or other viruses.

### **Aim 2: To map the expression of EBV Latency genes as quantitative traits**

Beyond the analysis of LMP1 eQTL, all the viral genes expressed in the latency modes will be quantified. eQTL mapping for each of the latency transcripts will be carried out using cDNA from two different sets of LCLs: the 58 Yoruban HapMap cell lines from the hypoxia study, and 352 British cell lines from a global eQTL study (MRC-A) (Moffat et al. 2007; Appendix Table A3). In the latter, differences in viral gene expression will be associated with human genetic markers and the outcome will be subsequently compared to the genome-wide human expression in the same LCL samples. The aim of this comparison will be to identify pairwise associations linking one SNP to both EBV latency transcripts and human transcripts. The same comparison will also be made for eQTLs from public databases – SCAN, Genevar and GTEx. Candidate eQTLs will then be prioritised depending on the strength of the association, presence of a distinct, continuous association peak and the overlap with human

transcript eQTLs. If successful, specific interactions between the viral and host proteins will subsequently be replicated in an independent experiment.

As an extension of the latency eQTL assay, the same method will be applied to public EBV expression data. EBV transcript expression data will be sourced from online RNA-seq experiment databases and used for GWA tests. Specifically, 5 RNA-seq studies in which EBV transcripts have been quantified will provide the phenotype data for the purpose of the current project (Arvey et al. 2012). All of these RNA-seq experiments were conducted using HapMap LCLs.

### **Aim 3: To map EBV copy number as a quantitative trait**

The next aim of the project will be to quantify EBV copy number across a large panel of over four hundred LCLs from the 1000 Genomes project. A target EBV genomic sequence will be amplified by qPCR and copy number will then be determined using a standard curve calibrated with serial dilutions of genomic DNA from a lymphoblastoid cell line such as Namalwa or Raji, which contain a fixed number of integrated viral copies per cell (MacMahon et al. 1991, Dehee et al. 2001). The relative copy number will then be used as a phenotypic trait in a genome-wide association test.

An extension of the experiment will involve mapping viral copy number QTLs across a panel of HapMap cell lines for which the relative EBV copy number has already been quantified by qPCR (Choy et al. 2008).

### **Aim 4: To perform an integrated analysis in order to determine the significance of results across all cohorts**

In order to understand the significance of the observed eQTLs and their relationship to reported disease associations, eQTL findings will be integrated with literature data and

reported GWAS associations (especially EBV-related), and prioritised based on their statistical and biological significance as well as replication in more than one cohort.

Viral expression and copy number experiments conducted using different cohorts will be subjected to a meta-analysis using GWAMA, a statistical tool developed at the Wellcome Trust Centre for Human Genetics (Magi and Morris 2010). The aim of the meta-analysis will be to compare the strength and the direction of identified regulatory effects across all available cohorts.

Additionally, correlation tests between EBV transcript expression levels and copy number will be conducted for a subset of HapMap LCLs for which both phenotypes are available. The same correlation test will also be conducted to check whether EBV transcript levels correlate with human transcript levels in the MRC-A cohort.

**Aim 5: To follow up and validate the most significant identified eQTL**

The final aim will be to follow up and replicate the top associations from the MRC-A cohort in primary human B cells and to generate LCLs from the same individuals in order to define functional effects of specific genetic variants.

## 2. Materials and Methods

### 2.1 Samples

This thesis studied a number of different LCL samples (Table 6).

Cohort ID	Source/Population	Aim of study	Publication	Individuals	Available for test
<b>Hypoxia</b>	HapMap / YRI	Latency eQTLs	Mohr 2010	58	58
<b>RNA-seq</b>	HapMap / YRI, CEPH	Latency eQTLs	Arvey et al. 2012 Cheung et al. 2010 Pickerell et al. 2010 Montgomery et al. 2010 <b>Public data</b>	67 YRI 67 CEPH	134 (67 + 67)
<b>MRC-A</b>	British nuclear families (asthma probands and sibs)	Latency eQTLs	Moffat et al. 2007	352	277
<b>1000G</b>	1000 Genomes / British, Finnish, Iberian, Italian	Copy number	Abecasis et al. 2012	414	287
<b>Choy</b>	HapMap / YRI, CEPH	Copy number	Choy et al. 2008 <b>Public data</b>	87 CEPH 85 YRI ( 45 JPT <u>meta-analysis only</u> )	172 (87 + 85)

**Table 6** – Summary of the LCL cohorts and populations tested in the current project. “Cohort ID” field lists the alternate name of each cohort adopted for the purpose of the study. “Source/Population” lists the origin of the samples and the ethnicity of the donors. “Publication” refers to previous studies conducted using the relevant sampleset. The total number of samples in a given cohort is listed in the “Individuals” field. The final field lists the actual number of samples available for testing (after QC) in the current study.

The LCLs were used to generate genomic DNA and RNA for analysis. In addition to the data provided by the current work, previously generated gene expression (for both human and EBV) and EBV copy number data was used in the study. This is specified separately for different cohorts as follows. LCLs were generated by the HapMap and 1000 Genomes Projects (cell lines and purified genomic DNA sourced from the Coriell Institute Cell Repository), and the MRC-A study (kindly provided by Professor William Cookson) (Moffatt 2007). EBV expression and copy number data from experiments by Choy et al. (2008) (indicated in Table 6 by “Choy” cohort ID) and RNA-seq studies by Cheung et al. 2012 et al. (2010), Pickerell et al. (2010) and Montgomery et al. (2010) was also used in the project (indicated by RNA-seq ID in Table 6).

### **2.1.1. Hypoxia study HapMap samples**

cDNA and RNA derived from Yoruban LCLs grown for the Mohr study (Mohr 2010) was used for LMP1 candidate eQTL validation and initial EBV latency eQTL mapping. These LCLs had been grown twice, in two independent experiments. The LCLs were sourced from 58 unrelated Yoruban individuals (HapMap Consortium) provided by Coriell Institute’s Cell Repositories and represented 2<sup>nd</sup>, 3<sup>rd</sup> and 4<sup>th</sup> passages of culture.

### **2.1.2. MRC-A**

RNA was sourced from 352 LCLs derived from sib pairs (Moffat et al. 2007). The MRC-A cohort consisted of nuclear families of British origin with at least one child diagnosed with asthma in each family (Moffat et al. 2007, Liang et al. 2013). All sibs had been genotyped using Illumina Sentrix HumanHap300 BeadChip (ILMN300K) or Illumina Sentrix Human-1 Genotyping BeadChip

(ILMN100K), or both, and a subset was subsequently imputed for HapMap Phase 3 genotypes (Liang et al. 2013). Global gene expression (approximately 50k transcripts) had been assayed across the whole cohort using Affymetrix HG-U133 Plus 2.0 chips (Liang et al. 2013). HapMap Phase 3 MACH-imputed genotypes had been kindly provided together with the global expression data by Professor Cookson. MRC-A samples were used to map viral eQTLs.

### **2.1.3 1000 Genomes**

Genomic DNA was sourced from 414 LCLs shipped from the Coriell Institute's Cell Repositories. The samples belonged to the 1000 Genomes Project Human Variation Panel and included 100 Finnish (Catalog ID: MGP00001), 100 British from England and Scotland (MGP00003), 114 Italians from Tuscany (MGP00007) and 100 Iberian individuals (MGP00010). The gDNA has been quantitated and normalised to a concentration of 100ng/ul using Qubit® 2.0 Fluorometer (Life Technologies). It was subsequently used in a copy number assay aimed to determine the EBV load in each individual LCL.

### **2.1.4 Namalwa LCL**

Genomic DNA from a Burkitt's lymphoma cell line with 2 integrated copies of the EBV genome per cell was obtained from Public Health England (Catalog number: 87060801). The gDNA concentration had been normalised to 100ng/ul (Nanodrop). The Namalwa gDNA serial dilutions were used to construct a standard curve for the EBV relative copy number assay.

### **2.1.5 Oxford Biobank**

Healthy volunteers of European ancestry were recalled by specific genotype from Oxford BioBank for the two most significant SNPs which showed evidence of both EBV and Human eQTL. Individuals were selected to include both major and minor allele homozygotes and heterozygotes for the informative SNPs. In total 39 volunteers with specific genotypes were recruited. These individuals provided blood for the purification of B cells and the establishment of LCLs (methods detailed in section 2.8). Whole blood samples from each individual were shipped in ACD tubes to Public Health England (Culture Collections' Genetic Support Services in Porton Down, Wiltshire UK) to be transformed with EBV and grown into LCLs. Frozen LCL culture samples were taken at 4 different time points throughout the outgrowth and cell pellets shipped back for RNA extraction and RT-PCR. This was done to compare gene expression at different time points.

## **2.2 RNA work (Oxford Biobank and MRC-A samples)**

### **2.2.1 RNA Isolation**

Qiagen RNeasy kits were used for RNA extraction from the purified B-cells from the Biobank samples according to the manufacturer's guidelines. Frozen cell pellets suspended in RNA buffer were thawed and homogenised on ice by pipetting repeatedly. Lysate was then applied to RNeasy mini columns in 2 ml collection tubes and subjected to RNA extraction. Briefly, ethanol was applied to allow binding of RNA to the columns and three washes were carried out before RNA was eluted into Eppendorf collection tubes with 50ul of RNase-free water and quantified with NanoDrop 1000 Spectrophotometer (Thermo Scientific).

### **2.2.2 Reverse Transcription (RT) PCR**

The Oxford Biobank and the MRC-A RNA samples were converted into cDNA for quantitative PCR. The Invitrogen Superscript III Reverse Transcriptase reaction kit (Catalog number 18080-085) was used for cDNA synthesis together with random hexamers (N8080127 Invitrogen) and Ribonuclease H from *E. coli* (AM2293 Invitrogen). Briefly, the initial reaction mixture of 1ul of RNA, 1ul of 50uM random hexamers, 1ul 10mM of dNTP mix, and 10ul of RNase free water was heated at 65°C for 5 minutes (to denature RNA), centrifuged and, after inclusion of 4ul of 5X First Strand Buffer, 1ul of 0.1 M DTT, 1ul of RNase OUT (40 U/ul) and 1ul of Superscript III RT (200 U/ul), incubated at 50°C for 50 minutes to achieve first complementary cDNA strand synthesis, then at 85°C for 5 minutes (to stop the RT enzyme). Finally, the RNA strand was degraded by adding 1ul of Rnase H (2U/ul) and heating the mixture for 20 minutes at 37°C. An MJ Research PTC240 Tetrad 2 thermal cycler was used for cDNA amplification. Reactions were performed in 96-well plates, and all reaction components were handled on ice. RNA and Superscript III enzyme were added at the end of each reaction. cDNA samples obtained were subsequently frozen and stored at -25°C as stocks. Additionally, 10ul of each sample was removed and diluted 10 times with RNase-free water to provide working solutions ready for q-PCR amplification.

### **2.3 Quantitative PCR (qPCR)**

Primer pairs targeting eleven EBV latency transcripts were sourced from the literature, from previously conducted experiments (Pan et al. 2006 Lindsey et al. 2009 Bell et al. 2006 Kelly et al. 2006). Primer specificity was subsequently tested across a range of annealing/extension temperatures (55-70°C) using an MJ Research PTC240 Tetrad 2 thermal cycler.

### 2.3.1. SYBR Green Quantitative Real Time PCR

SYBR Green qPCR method was used to quantify the latency gene transcripts in the sibling panel from the MRC-A cohort. It relies on a chemical dye that emits fluorescent green light when bound to double-stranded nucleic acids. Levels of fluorescence are proportional to the levels of the target transcript amplified by the primers. As recommended by the manufacturer's instructions, 6.25ul of iQ SYBR Green Supermix (2x concentration, which includes *Taq* polymerase, SYBR fluorescent dye, fluorescein, MgCl<sub>2</sub>, dNTPs and stabilisers; Catalog number 170-8880) was mixed with 1ul of primers, 1ul of cDNA template and 4.25ul of RNase free water to make up 12.5ul total volume reactions, which were then run on the CFX BioRad cycler. Reactions were performed in 96-well plates, and all reaction components were handled on ice. Bio-Rad adhesive transparent seals (Microseal B, adhesive, optical; Catalog number MSB-1001) were applied and the plates centrifuged at 500rpm for 1 minute before each run. The cycler was set to SYBR Green I dye, 12ul reaction volume, and the cycling conditions were: 95°C for 8 minutes to activate the polymerase enzyme, followed by 40 cycles at 95°C for 10 seconds (to denature the strands) and 20 seconds at 60°C (for annealing and strand extension), followed by 20 seconds at 72°C. Melting curve analysis was also performed at the end of each plate run to check for non-specific amplification. The conditions were 10 seconds at 95°C , 5 seconds at 65°C and 5 seconds at 95°C. For all latency transcripts, reactions were conducted in duplicates. Raw Ct values were calculated and obtained using Bio-Rad CFX software (v2.0). Wherever the standard deviation (SD) exceeded 1,5 repeat reactions were conducted in triplicates. The two most similar replicate values were retained (if SD was below 1). All Ct values were then averaged and normalised according to two riboprotein housekeeper genes adopted for the assay (which were also run in duplicates on each same plate).

### 2.3.2 TaqMan Quantitative Real Time PCR

TaqMan qPCR assays were performed using the Applied Biosystems ABI7900 machine in the 'Standard' mode. Pre-designed assays were used with Applied Biosystems TAMRA fluorescent probes. To determine EBV copy number in the 1000 Genomes Human Variation panel samples, a relative copy number quantification was performed using EBV *IR1* gene (Applied Biosystems catalog number: 4331182, assay code Pa03453399\_s1) and RNase P (Applied Biosystems, Catalog number 4316831) TaqMan assays. 20ul reactions were prepared including 1ul of gene-specific 20x TaqMan Gene Expression assay, 1ul of sample gDNA solution, 10ul of 2xTaqMan Universal PCR Master Mix and 8ul of RNase free water. The cycling conditions were set to 50°C for 2 minutes, 95°C for 10 minutes, followed by 40 cycles at 95°C for 15 seconds and 60°C for 1 minute. Also the Namalwa cell line was used to estimate the absolute concentration / copy number of EBV. Since the Namalwa gDNA has been normalised to a uniform concentration of 100 ng/ul and each Namalwa cell contains 2 copies of EBV haploid genome integrated into the cellular genome, it was possible to ascertain the unknown DNA concentration and copy number in the 1000 Genomes Project samples (whose concentration has also been normalised to 100 ng/ul). The required information was estimated from a standard curve constructed using serial dilutions of the Namalwa DNA. Copy number concentration was inferred by dividing the weight of DNA by 6.6 pg (which is the weight of a diploid cellular genome). Reactions were set using Applied Biosystems MicroAmp® Fast Optical 96-Well Reaction Plates (Catalog number 4346906) sealed with Applied Biosystems MicroAmp® Optical Adhesive film (Catalog number 4360954). Plates and all reaction components were handled on ice.

Expression of the following genes, flanking the previously identified LMP1 eQTL (Mohr 2010) was quantified using custom TaqMan probes: *CMY1* (*XIRP1*) (ABI assay code

Hs01589203\_m1), *CX3CRI* (ABI code Hs01922583\_s1), *WDR48* (ABI code Hs00368247\_m1) and *CCR8* (ABI code Hs00174764\_m1). Results were normalised to *OAZ1* (ABI code Hs00427923\_m1), the housekeeping gene which was used in the original study. TaqMan Fast Universal PCR Master Mix (Catalog number 4352042) was used, and the ABI7900 machine run in the 'Fast' mode with reaction volumes reduced to 10ul.

## **2.4 Housekeeper Assay**

Transcript quantification has to overcome the caveat of differing starting amount of RNA that in turn affects the levels of measured transcript, differential polymerase efficiency as well as inter-individual or tissue-specific variability in gene expression. Accurate normalisation is therefore necessary to compare expression in different individuals (Vandermompele 2002, Pfaffl 2004). The amount of transcript can be standardised to the number of cells the RNA was extracted from, or the total RNA mass, however internal control genes are the most frequent method applied in expression normalisation (Vandermompele 2002). These are genes thought to be stably expressed across different individuals, tissue types and conditions (Pfaffl 2004). However, evidence from literature indicates that there is certain inherent variation and sometimes housekeeper genes may not be stably expressed, therefore checking their stability prior to an experiment as well as normalisation to multiple housekeeper genes have both been recommended to ensure a better quantification of target transcripts (Vandermompele 2002, Pfaffl 2004).

### **2.4.1 BestKeeper**

In order to select stable housekeepers suitable for internal normalisation of the Ct values obtained by assaying the MRC-A samples for latent EBV expression, BestKeeper and geNorm programmes have been used. BestKeeper is a Microsoft Office Excel based tool developed by

Pfaffl et al. (2004) and designed to investigate expression stability of multiple (up to ten) housekeeper genes at one time and across independent technical and biological replicates. The input for Bestkeeper is the threshold cycles (*Ct*) or crossing points (CP) averaged for each sample across all replicates. *Ct* units are usually normally distributed and usage of *Ct* units is comparable with a logarithmic data transformation and enables parametric statistical tests to be used (Pfaffl 2004). From this data standard descriptive statistics are calculated including SD and a coefficient of variance. Genes with high SD exceeding 1 are eliminated from further analysis. The software then calculates and relies on a function called the BestKeeper Index which is a geometric mean of the average *Ct* (or CP) values of all housekeeper genes which are put to test, and is given by the formula:

$$z \sqrt{(CP1 \times CP2 \times CP3 \times \dots \times CPz)}$$

Descriptive statistics for the BestKeeper index are also calculated. The programme then takes all of the input housekeeper genes' average *Ct* values and makes pairwise correlation tests between each two and between each gene and the combined BestKeeper index, calculating the P-value and the Pearson correlation coefficient (*r*). The relation of each gene and its contribution to the index is given by the correlation coefficient (*r*) and the coefficient of determination (*r*<sup>2</sup>). This determines which of the contributing genes correlate most strongly with the index (and with each other). Genes whose expression is highly correlated (>0.9) and which produce a consistent BestKeeper Index are then retained for the normalisation of other target genes *Ct* values. Alternatively, the BestKeeper Index itself can be used as a stable standardising index.

### 2.4.2 geNorm

A similar Excel-based tool developed by Vandesompele et al. (2002) also accepts Ct values as input and assumes that ideal housekeepers have always the same stable ratio of expression levels, regardless of other factors. For each tested candidate gene, geNorm computes a statistic called the internal control gene stability measure  $M$ . First, pairwise variability is calculated. For each two-gene combination ( $j$  and  $k$ ), an array of logarithmically transformed expression ratios ( $a_j$  and  $a_k$ ) is established. Log-transformed ratios follow a normal distribution centered on zero and take on equal absolute values with opposite signs for any specific ratio and its inverse (Vandesompele et al. 2002). The length of the array depends on the number ( $m$ ) of samples or replicates:

$$A_{jk} = \left\{ \log_2 \left( \frac{a_{1j}}{a_{1k}} \right), \log_2 \left( \frac{a_{2j}}{a_{2k}} \right), \dots, \log_2 \left( \frac{a_{mj}}{a_{mk}} \right) \right\} = \left\{ \log_2 \left( \frac{a_{ij}}{a_{ik}} \right) \right\}_{i=1 \rightarrow m}$$

The stability measure  $M$  is calculated as the arithmetic mean of SDs of all arrays that combine a particular gene with all other control genes. Genes with the lowest  $M$  values have the most stable expression. This ranks the genes according to the expression stability and stepwise exclusion of housekeepers with highest  $M$  value is carried out until the two most stable remain within the index (Vandesompele et al. 2002).

### 2.5. eQTL Mapping

eQTL analysis is a powerful approach to evaluate regulatory genetic variants by finding correlations between gene expression and genotypes. With the advance of high-throughput technologies, large datasets make eQTL analysis technically challenging. There are multiple

approaches to eQTL studies, however most involve multiple independent association tests for each transcript-SNP pair, using linear regression and ANOVA models (PLINK) and non-linear methods such as mixed linear models and Bayesian regression (Shabalin 2012).

### **2.5.1 Plink**

PLINK (Purcell et al. 2007) was used to map EBV viral copy number and latency eQTLs in all cohorts composed of unrelated individuals. Plink is a C/C++ statistical GWAS tool set designed to work with high-throughput data (such as genotype files containing hundreds of thousands of genotyped markers in thousands of individuals) and capable of finding SNPs that correlate with the amount of transcript through association tests for quantitative traits employing a standard linear regression of phenotype on allele dosage. Plink relies on \*.ped files (containing information about individuals and their genotypes as well as phenotypes, which can also be specified in a separate \*.pheno file) and \*.map files (with information on the chromosomal location of all the genotyped SNPs) – formats widely employed in GWAS. To comply with Plink input format specifications, expression and viral copy number phenotypic values have been used in a separate \*.pheno file, while genotypes for all individuals from HapMap and 1000 Genomes panels have been downloaded as \*.ped and \*.map files, and \*.vcf files respectively, from the HapMap and 1000 Genomes project websites. VCFtools were subsequently used to convert the relevant files into \*.ped and \*.map format. All files were transformed into binary files in PLINK in order to shorten the time necessary for association testing.

PLINK tests quantitative traits for association using linear regression and asymptotic significance P-values (Wald test) (Holm et al. 2010, Purcell et al. 2007). Wald test is a popular test of linear restrictions on the coefficients of the standard linear regression (Lo et al.

1985). Originally, the  $W$  statistic had a quadratic form containing the estimated information matrix and the linear restrictions (Lo et al. 1985). The command `--assoc` was used to carry out the association test twice, with 0.01 and 0.05 minor allele frequency filter thresholds, to reduce false positive findings. When presented with a quantitative phenotype like copy number or transcript level (provided in a `*.ped` or `*.pheno` file), the `--assoc` command computes and produces regression statistics and Wald test results and saves them into a `plink.qassoc` file. This is equivalent to a 2 degrees of freedom genotypic association test.

`*.qassoc` files were used to obtain global association Manhattan plots with a short script written in R. The SNP and the P-value column data was also used to construct local association plots for the most significant association peaks observed using an online tool LocusZoom. For the most interesting associations, relevant genotypes of all tested individuals were sourced from the `*.ped` file to construct box plots showing changes in gene expression and their direction dependent on the genotype.

### **2.5.2 Merlin-offline**

To test the MRC-A family cohort composed mostly of sib pairs with HapMap Phase 3 MACH-imputed genotypes, a different approach was employed. MACH is a Markov Chain based haplotyper able to infer missing genotypes in samples of unrelated individuals created by Yun Li and Goncalo Abecasis (<http://www.sph.umich.edu/csg/abecasis/MACH/index.html>). Merlin-offline was used because of its ability to both work with imputed genotype files and perform family based association tests. The current version is a pre-release. Merlin-offline is an undocumented programme created by Yun Li and Goncalo Abecasis and distributed together with the statistical computer software MERLIN created by Yun Li et al. (2010). Merlin-offline can work with Mach/MiniMac and other imputed genotype files once they are

converted to MERLIN \*.ped, \*.dat, \*.fam and \*.map format. Merlin-offline can work with fully genotyped cohorts, however it can also estimate missing genotypes using probabilistic Lander-Green or Elston-Stewart (for large pedigrees with over 15 individuals per family) algorithms (Chen and Abecasis 2006). Merlin-offline then evaluates genome-wide associations for quantitative traits on the imputed dosage scores under an additive model through a Mixed Linear Model score test that incorporates an expected or pre-specified identity-by-descent (IBD) kinship matrix (Kampert 2011, Benyamin et al. 2009, Martin et al. 2011, Chen and Abecasis 2006). The score test is normally specified by MERLIN's --FastAssoc function. Alternatively, a likelihood-ratio test (--assoc function) can be applied, however the authors recommend the score test as faster and more suitable for "screening very large numbers of markers" like in GWAS, while reserving the likelihood-ratio test for a more accurate analysis of smaller marker sets typical of regional association tests in follow-ups and replication experiments (MERLIN Tutorial website). Since the experimental cohort included over 300 individuals in small pedigrees and over 2 million markers, the two tests should, according to authors, produce very similar results (MERLIN Tutorial website). Consequently, the faster score test was applied. The relevant HapMap Phase 3 imputed genotypes and family information files were obtained from collaborators in the MERLIN-compatible format. Association test results were saved into merlin-fastassoc-chr\*.tbl files with the following fields: Marker / Allele / Effect / H2 / LOD / P-value. As with PLINK, the Marker and P-value data was subsequently used to construct global and regional association plots.

### **2.5.3 Matrix eQTL**

To analyse the expression data from the RNA-seq experiments that used Caucasian and Yoruban HapMap samples, Matrix eQTL programme was used. Matrix eQTL is the official statistical eQTL analysis software of the Genotype-Tissue Expression project (GTEx) project

that aims to construct “a public atlas of gene expression and regulation across multiple human tissues” (Shabalín 2012, Lonsdale et al. 2013, GTEx Portal). The reason for this was the large phenotype set containing expression values for 83 EBV lytic and latency transcripts. Matrix eQTL has an inbuilt feature allowing it to carry out a principal components analysis (PCA) of the multiple phenotypes, and then apply a chosen number of principal components as covariates in the association test (in order to control for confounding by population structure or technical variation). The software performs linear regression (least squares) analysis using the specified set of covariates and calculates a t-test statistic for every transcript and SNP, and then, to make the analysis faster, computes P-value, only for the test statistics that exceed a certain significance threshold, specified by the user (Shabalín 2012).

## **2.6 eQTL data integration**

In order to prioritise association findings and discover possible functional links to human cellular pathways, public human eQTL databases were searched for those eQTLs that overlap with EBV QTLs spatially within a 200kb window. Before mining public study data and in order to obtain additional SNPs remaining in high LD with those already genotyped and tested, an LD-based proxy-search was carried out using an online tool, SNAP available at the Broad Institute’s website (standard  $r^2$  threshold 0.8, distance limit 500 and 1000 Genomes Pilot 1 settings applied). eQTL databases searched included SCAN, Genevar, GTEx, The eQTL Browser (Pickerell et al. 2010) and seeQTL (Xia et al. 2011).

## **2.7 GWAMA**

Meta-analysis combining identical phenotypes across multiple cohorts used in the study was conducted using Genome-Wide Association Meta-Analysis (GWAMA), a statistical software developed at the Wellcome Trust Centre for Human Genetics (Magi and Morris 2010). Fixed-

effect analysis (which assumes that alleles of a SNP have the same effect on a quantitative trait in each tested populations) was performed for EBV transcript phenotypes and EBV copy number. GWAMA accepts PLINK format output files using the summary statistics to perform meta-analysis. It checks the alignment of the genotyped SNPs onto same reference strand. Fixed effects meta-analysis looks at each SNP effect by combining its allelic effects weighted by the inverse of their variance. The combined allelic effect of the  $j$ th SNP is calculated with the equation:

$$B_j = \frac{\sum_{i=1}^N \beta_{ij} w_{ij}}{\sum_{i=1}^N w_{ij}},$$

Where  $\beta_{ij}$  is the strand-aligned effect of the reference allele at the  $j$ th SNP in the  $i$ th study;  $w_{ij} = [\text{Var}(\beta_{ij})]^{-1}$  is the inverse of the variance of the estimated allelic effect in the  $i$ th study, obtained from the standard error. GWAMA also checks the direction of each SNP's effects in all tested cohorts and reports the consistency as well as checks the summary statistics to control for population structure by reporting the lambda genomic inflation control factor (Devlin and Roeder 1999). This is a factor given by the median of the test statistics, divided by its expectation under the null hypothesis of no association, which is used to multiply and uniformly correct the test statistics inflated due to the population structure (Magi and Morris 2010).

## **2.8. Oxford Biobank Whole Blood preparation**

### **2.8.1 Sample processing**

Whole blood samples were obtained from healthy volunteers of specific genotypes from the Oxford Biobank in ACD and EDTA tubes. The ACD tubes were subsequently shipped on the same day to Public Health England Cell Culture services for EBV transformation and LCL outgrowth, while the EDTA tubes were retained for B-cell purification.

### **2.8.2. B-cell sorting**

7 EDTA tubes per individual, containing approximately 60ml of whole blood were obtained from the Oxford Biobank. The blood from each individual was pooled into two 50 ml Greiner tubes containing 20 ml of HBSS buffer to a total volume of approximately 50 ml per tube. The HBSS-diluted blood samples were subsequently transferred from the first set of tubes into another set for Ficoll extraction. For each two 50 ml Greiner tubes with HBSS-diluted blood from a single individual, three 50 ml Greiner tubes containing 12.5 ml of Ficoll were prepared. The blood was then transferred carefully, by tilting the tube and gently pouring the blood down the side of the Ficoll tube so that a layer of blood formed on top of Ficoll solution. Once the samples were transferred into the new set of tubes, they were centrifuged for 30 minutes at 400g and room temperature with no brake (in order not to disturb the blood layer).

In the next step samples underwent washing in HBSS solution. After spinning, the topmost plasma layer was removed using an automated pipette, and a sterile Pasteur pipette was used to remove the PBMC layer, which was pooled for each individual in a new set of 50ml Greiner tubes (single tube per individual) filled with 30 ml HBSS pre-cooled on ice. The samples were centrifuged for 10 minutes at 300g, room temperature. Supernatant was discarded. Pellets were

resuspended in 30 ml of HBSS and subjected to another spin, then, after removing the supernatant, resuspended again in 20 ml DPBS solution (0.5% Fetal Bovine Serum). Cells were counted and the samples pelleted. In the next step, supernatant was removed and, depending on the cell count in each sample, the pellets were resuspended in an adequate volume (as given by the Miltenyi Biotec CD19 MicroBead protocol) of PBS buffer (0.5% Fetal Bovine Serum; 80ul per 10 million PBMCs) and magnetic CD19 Bead solution (20ul per 10 million PBMCs). The solution was left to incubate for 15 minutes at 4°C and was centrifuged for the fourth time. Supernatant was discarded and pellets resuspended in 0.5% Fetal Bovine Serum PBS solution – either 0.5 ml if sample counted up to 100 million PMBCs, or 1 ml if it numbered more cells. The solution was then fed into the Miltenyi AutoMACS machine, and B-cell counts were obtained from a 10ul fraction of the 2ml effluent. Approximately one third of the effluent was then transferred into a set of eppendorfs and frozen in 300ul of RLT buffer. The remaining solution was centrifuged in falcon tubes, resuspended in 20% Fetal Bovine Serum RPMI solution (0.5ml if up to 1mln B-cells, or 1ml if more), split in two and transferred into two sets (one to be cytokine-stimulated, and one control) of 5ml tubes ready for incubation.

### **2.8.3. Stimulation and harvesting of B cells**

B-cells were left to recover at 37°C, 5% CO<sub>2</sub> overnight, after which one set of samples was stimulated with CD40L (2ug/ml, 1ul in 500ul) CpG ODN (0.5uM, 1ul in 500ul) and IL-4 (20ng/ml, 5ul in 500ul) and incubated for 24 hours. The other one was also incubated, but with no stimulation. The cells were then harvested by centrifugation at 300g and 4°C for 5 minutes, the supernatant was discarded and pellets re-suspended in 300ul RLT and stored at -80°C.

## Chapter 3. Mapping EBV transcript QTLs

### 3.1 Introduction

Previous work has shown that human genetic variation was associated with differences in the expression of the EBV latency gene LMP1 in LCLs (Mohr 2010). The most significant association to LMP1 expression was rs1913243 (P-value  $5.8 \times 10^{-8}$ ). One hypothesis for the mechanism driving this association is that rs1913243 or other variants in strong LD act in *cis* to modulate the closest genes and somehow affect the process of viral infection/activity such that differences in expression of LMP1 occur. Expression levels of the two closest flanking genes to rs1913243, *CMYAI* and *CX3CRI*, were not quantified at the time of the original study (Mohr 2010). This chapter describes work to test this hypothesis together with the results of eQTL mapping of EBV latency transcripts in different LCL cohorts (cohorts detailed in Table 6 and section 2.1).

### 3.2 Aims of the chapter

1. To test the hypothesis that the candidate eQTL (associated with EBV LMP1 transcript level), rs1913243, is also associated with expression of the nearest flanking human genes (as a potential *cis*-acting eQTL) in the YRI HapMap LCLs used by Mohr and colleagues (Mohr 2010).
2. To extend the eQTL analysis in YRI HapMap LCLs by quantification and mapping of other EBV latency eQTLs.
3. To quantify and map EBV latency transcript eQTLs in a larger, independent cohort of LCLs from the MRC-A panel established by Moffatt and colleagues (Moffatt 2007).

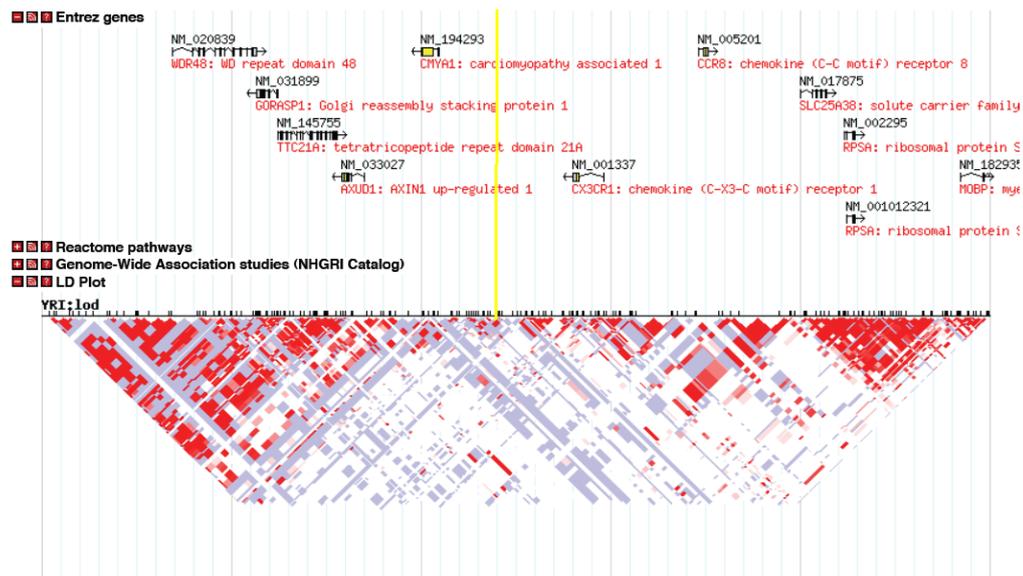
4. To map EBV latency eQTLs using EBV transcript expression levels available from RNA-seq experiments conducted on HapMap LCLs and compiled by Avery and colleagues (Avery 2013).
5. To integrate MRC-A EBV latency eQTLs with MRC-A human eQTLs and prioritise candidate SNPs showing association with both viral and human gene expression.

Cohort ID	Source/Population	Aim of study	Publication	Individuals	Available for test
<b>Hypoxia</b>	HapMap / YRI	Latency eQTLs	Mohr 2010	58	58
<b>RNA-seq</b>	HapMap / YRI, CEPH	Latency eQTLs	Arvey et al. 2012 Cheung et al. 2010 Pickerell et al. 2010 Montgomery et al. 2010 <b>Public data</b>	67 YRI 67 CEPH	134 (67 + 67)
<b>MRC-A</b>	British nuclear families (asthma probands and sibs)	Latency eQTLs	Moffat et al. 2007	352	277
<b>1000G</b>	1000 Genomes / British, Finnish, Iberian, Italian	Copy number	Abecasis et al. 2012	414	287
<b>Choy</b>	HapMap / YRI, CEPH	Copy number	Choy et al. 2008 <b>Public data</b>	87 CEPH 85 YRI ( 45 JPT <u>meta-analysis only</u> )	172 (87 + 85)

Table 7 – Summary of cohorts and samples used in the current study.

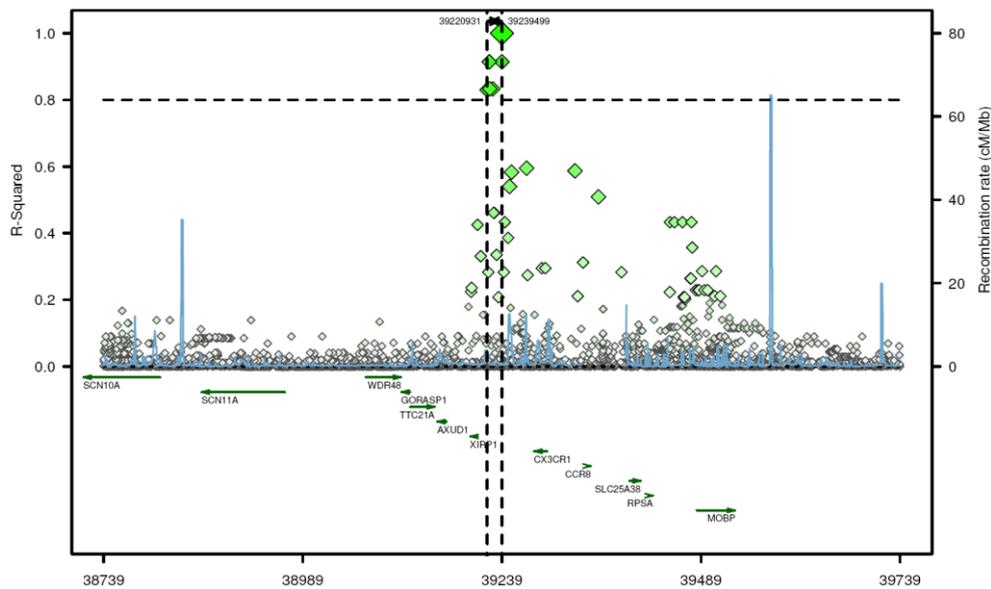
### 3.3 Characterisation of *cis* effects of the candidate LMP1 eQTL, rs1913243

Haplotype analysis using HapMap Genome Browser release #2 showed that in the Yoruban HapMap population, the region harbouring LMP1-associated eQTL is split into two to three major linkage disequilibrium blocks, with rs1913243 (the eQTL SNP) located in the middle (Figure 18).



**Figure 18** – Genomic landscape and LD plot spanning rs1913234, with names and locations of the neighbouring genes. Figure prepared using HapMap Genome Browser website (rs1913243 eQTL shown highlighted in yellow). LOD < 2, D' < 1 plotted in white; LOD < 2, D' = 1 in blue; LOD > 2 D' < 1 in shades of pink/red; and LOD > 2, D' = 1 in bright red.

A more localised analysis of LD structure using local LD-plots constructed with an online tool, SNAP, revealed that rs1913243 tended to remain more closely in linkage with the SNPs from the downstream LD block (Figure 19).

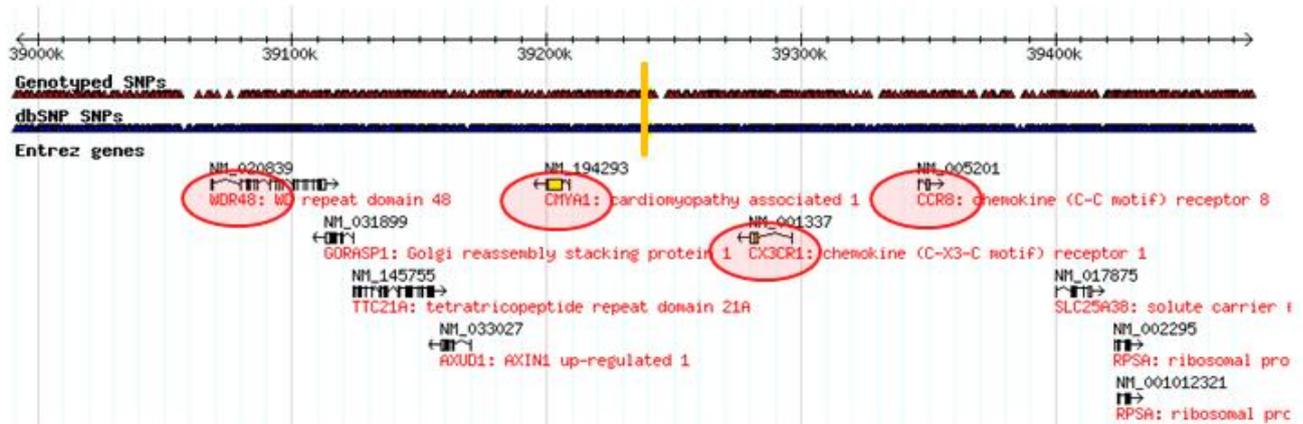


**Figure 19** - LD Plot constructed using SNAP online tool showing rs1913243 eQTL (large green diamond at the top) and the tested region. All 626 tested SNPs are displayed as smaller diamonds. LD between rs1913243 is indicated on the Y-axis (R-squared) as well as by the intensity of the shades of green and the size of the diamonds. The locations and names of genes located within the tested region are shown at the bottom and genomic location in kilobases is indicated by the X-axis.

A number of candidate genes located within a 485 kb window (upper margin extent of *cis* effects, Cookson et al. 2009) were considered and four genes (Figure 20) selected for transcript quantification and analysis for the following reasons:

- *CMYA1 (XIRP1)*: proximal to the putative eQTL, involved in cytoskeleton organization and negative regulation of cell proliferation.
- *CX3CR1*: co-receptor of HIV-1, variants may increase susceptibility to HIV infection (Faure et al. 2003 Combadiere et al. 2003).
- *WDR48*: may play a role in Herpesvirus saimiri and Human Papillomavirus HPV11 infection, regulator of de-ubiquitinating complexes (Ulrich and Walden 2010 Aviel et al. 2000 Jang et al. 2005), strong *cis*-acting eQTL present (rs205613 P  $2.8 \times 10^{-11}$ ) (Lunemann et al. 2009).

- *CCR8*: co-receptor of HIV-1; EBV, HHV8 and MCV express *CCR8* agonists (Lee et al. 2000 Efremov et al. 1999 Jinno et al. 1998 Horuk et al. 1998 Ruibal-Ares et al. 2004 Luttichau et al. 2010).



**Figure 20** – HapMap Genome browser diagram showing the genes (marked by red circles) selected for validation of the LMP1 eQTL (rs1913243, marked by a yellow bar). HapMap Genome Browser release #28 - Phases 1, 2 & 3 - merged genotypes.

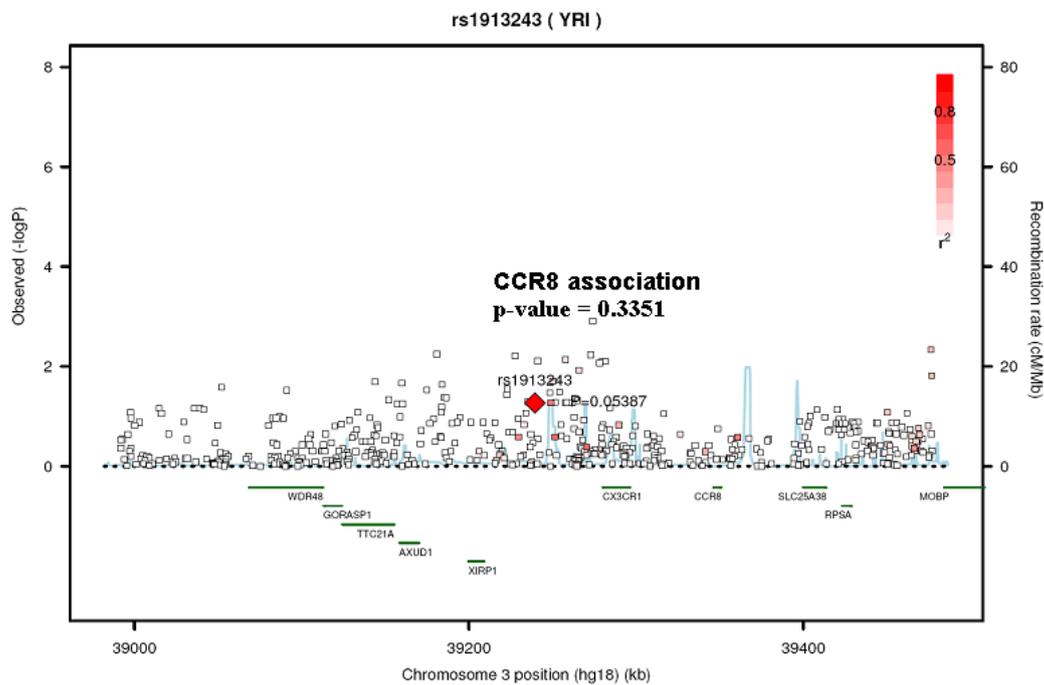
Potential *cis* regulatory effects of the LMP1 eQTL involving four genes in the flanking regions were investigated using exon-spanning TaqMan Gene Expression probes run on the 7900HT Fast Real-Time PCR System. cDNA from 58 LCLs from the Yoruba HapMap panel (listed in Appendix , Table A1) prepared for the previous study (Mohr 2010) was used. The cDNA samples included material sourced from 58 LCLs grown twice under different conditions: A) hypoxia-stimulated LCLs, B) cell lines grown in normoxic conditions. Both sample sets were assayed. For each set, for each cell line qPCR reactions were conducted in triplicates. Expression was normalised to a housekeeper *OAZ1* encoding ornithine decarboxylase antizyme 1. *OAZ1* was the most stably expressed housekeeping across the 58 LCLs when tested by the SLqPCR R package (geNorm) together with 6 other candidates (Mohr 2010). Regional association tests were conducted using Plink (MAF >5%, HWE  $p < 0.001$ , missingness 0.1) within an approximately 500 kb fragment (bp 38,991,792 to

39,477,047; fragment length was 485,255 bp in total), roughly 250kb on each side of the candidate eQTL, which is an approximate extent of potential *cis* regulatory effects. The selected fragment contained 626 SNPs (HapMap Phase 3). This yielded weak associations, all of which were considered non-significant after the Bonferroni multiple test correction for 626 independent tests (and  $2 \times 626$  when correcting for each trait being tested twice in Normoxia and Hypoxia samples). The results were also insignificant when Bonferroni Step-down (Holm) correction and FDR was considered instead. Even though the 4 tested traits may be independent no further correction was applied. The five most significant SNPs for each assay are listed in Table 8.

<b>CCR8 – Normoxia:</b>		<b>CX3CR1 Normoxia:</b>	
<b>SNP</b>	<b>P-Value</b>	<b>SNP</b>	<b>P-Value</b>
rs813670	0.05387	rs1917419	0.0007
rs784519	0.05387	rs6807342	0.00285
rs784510	0.05477	rs12497322	0.003797
<b>CCR8 – Hypoxia:</b>		<b>CX3CR1 Hypoxia:</b>	
<b>SNP</b>	<b>P-Value</b>	<b>SNP</b>	<b>P-Value</b>
rs813670	0.003344	rs11707681	0.000522
rs784519	0.003344	rs3755592	0.002687
rs784510	0.003437	rs4676627	0.01014
<b>XIRP1 – Normoxia:</b>		<b>WDR48 Normoxia:</b>	
<b>SNP</b>	<b>P-Value</b>	<b>SNP</b>	<b>P-Value</b>
rs9809888	0.0005	rs4676483	0.002078
rs9837442	0.003597	rs7651447	0.0099
rs9830167	0.005114	rs7621701	0.0122
<b>XIRP1 – Hypoxia:</b>		<b>WDR48 Hypoxia:</b>	
<b>SNP</b>	<b>P-Value</b>	<b>SNP</b>	<b>P-Value</b>
rs9809888	0.006845	rs2088667	0.000313
rs9837442	0.01057	rs17038663	0.002845
rs17791786	0.01701	rs4676483	0.005117

**Table 8** – Summary of results. Three most significantly associated SNPs (out of the 626 tested) are listed for each of the tested transcripts in both Normoxia- and Hypoxia-derived samples.

Local association plots are shown, constructed using the SNAP tool for *CCR8* (Figure 21) and other genes in Appendix Figures A1-A3. In conclusion, the LMP1 eQTL SNP rs1913243 was not associated with expression levels of the four flanking genes tested at the required level of significance.



**Figure 21** – Regional association plot of the tested sequence containing LMP1 candidate eQTL, rs 1913243. P-value of the rs1913243 association and the name of the tested transcript are indicated in bold within the plot. Genomic location is shown by the X-axis and P-value displayed on the Y-axis. 626 tested SNPs are depicted as squares; rs1913243 as the red diamond. SNPs in LD with rs1913243 are coloured in shades of red according to the extent of LD given by r-squared scale in to top left corner. Local genes are displayed at the bottom and their locations indicated by green lines. The plots shows association results obtained from the normoxia samples only.

### 3.4 EBV latency eQTLs in the Hypoxia study LCLs

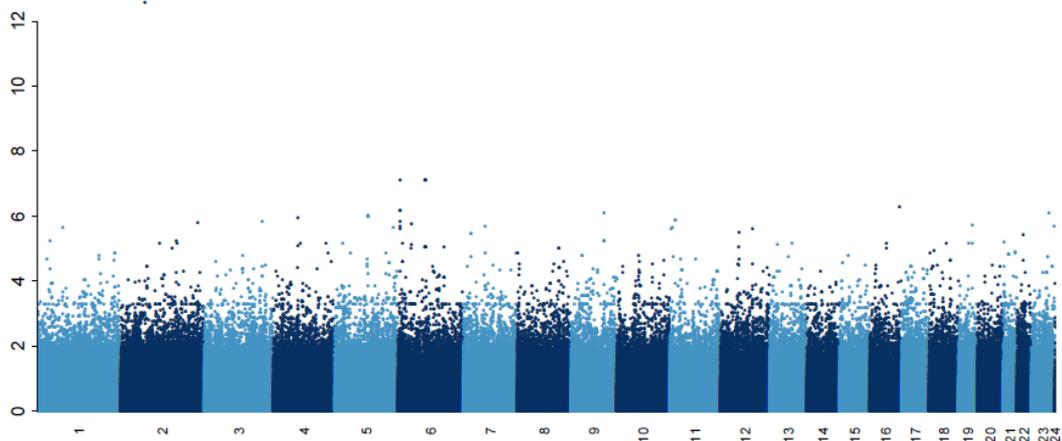
Since all EBV latency transcripts are important for the *in vivo* immortalisation process, and each family of latency proteins affects different molecular pathways and interacts with variable sets of human proteins, latency transcript abundance may also respond to different

regulatory genetic effects. Polymorphisms in human TFs should thus affect viral gene expression. Quantification of all latency transcripts was therefore the next aim of the study. It was expected that the key viral oncogenes, LMP1 and EBNA2, with the highest number of known molecular interactions, would be most sensitive to genetic regulation.

The first stage of the experiment was conducted using cDNA samples from the Mohr study. This also enabled revalidation of the candidate LMP1 eQTL using genetic material from the same cell lines that had been grown for the purpose of the hypoxia experiment. Consequently, cDNA from the 58 Yoruba LCLs, grown in normoxic conditions, was assayed for the expression of 11 EBV latency transcripts using BioRad SybrGreen quantitative PCR with primers that had been validated in previous studies (Pan et al. 2006 Lindsey et al. 2009 Bell et al. 2006 Kelly et al. 2006). The housekeeping gene *OAZ1* was used to normalise the expression values. Reactions were conducted in duplicates and, if SD of the replicates exceeded 1, repeated in triplicates, and averaged. Whole genome association tests were conducted using Plink (--assoc function) and HapMap phase 3 genotype data (28<sup>th</sup> May 2010 release, included 1,457,897 SNPs). Filters included 10% for individual and SNP missing rate, P-val HWE<0.001 and MAF > 0.05. The associations for different EBV latency transcripts are summarised below.  $5 \times 10^{-8}$ , a standard genome-wide significance threshold adopted in GWAS studies, was set as the P-value threshold for associations. Since viral latency phenotypes are expected to be highly correlated (most transcripts are produced from a common precursor transcript) they cannot be treated as genuinely independent observations. Consequently no further correction was applied particularly the Bonferroni correction, since it would be inaccurate when multiple phenotype traits are correlated.

### 3.4.1. BART

BARTs are untranslated micro-RNAs from the viral BamHI region whose function has not been established although they are expressed in all latency modes and appear essential for viral persistence (Young and Rickinson 2004, Kieff et al. 2010). A suggestive ( $1E^{-07} < P\text{-value} < 1E^{-05}$ ) association was found on chromosome 6, close to the centromere (Figure 22). The SNPs are in a segment with no gene located within 250 kb on either side. Another suggestive association to a single SNP was located on the same chromosome, in an intergenic region between *SLC22A23* (implicated in Crohn's disease) and *C6orf145* (Franke et al. 2010). A single significantly associated SNP, rs1380703, was located in an intergenic region on chromosome 2 (P-value  $2.70E^{-13}$ ).

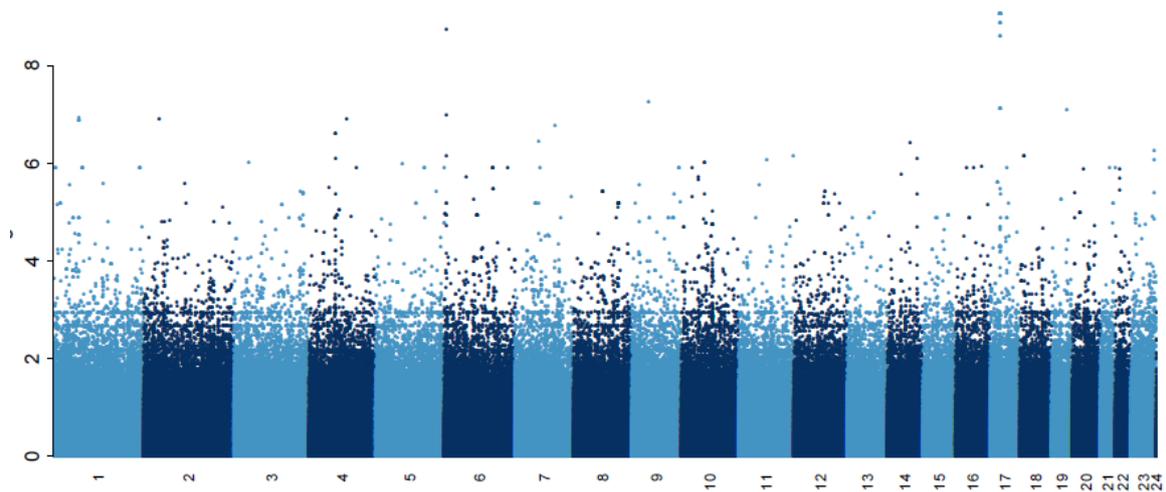


**Figure 22** – Whole genome association plot for BART transcript. Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10}P\text{-values}$  displayed on the Y axis.

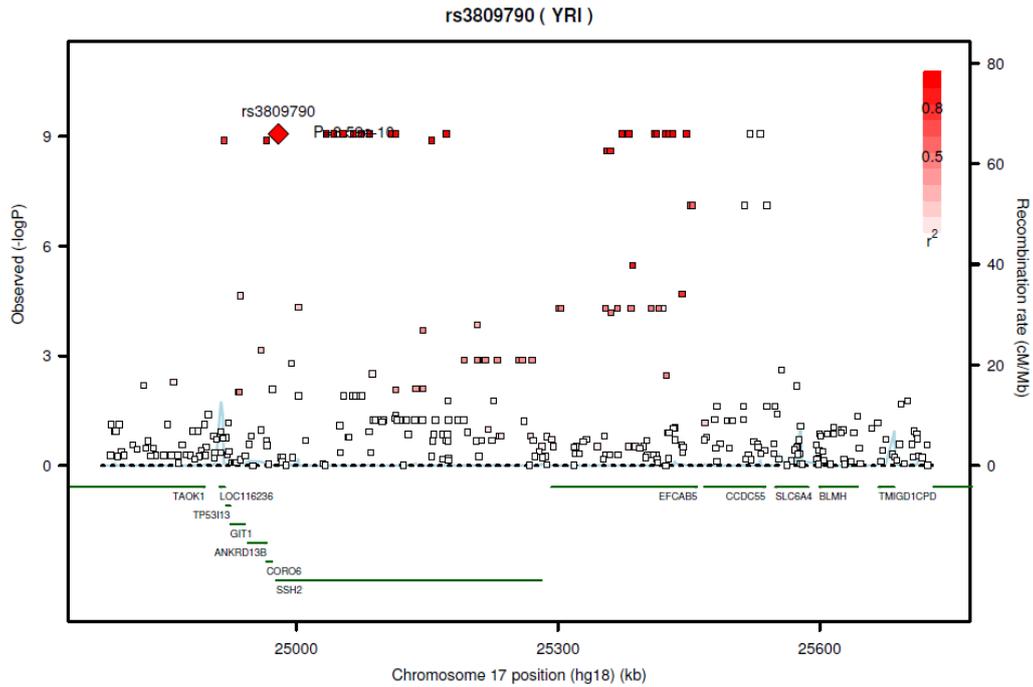
### 3.4.2 EBER1

EBER-1 and EBER-2, together with cellular factors, assemble into a ribonucleoprotein and inhibit kinase protein kinase R which mediates the antiviral effects of interferons. EBER-mediated inhibition of protein kinase R could promote viral persistence (Kieff et al. 2010)

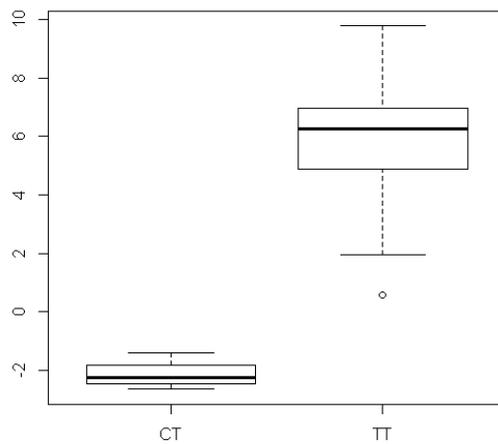
Young and Rickinson 2004). The SNPs most significantly associated with EBER1 expression cluster in one major site on chromosome 17 approximately 500 kb long (P-value  $8.59E^{-10}$ , Figure 23-25). Three genes can be located within the site, including *SSH2*, *FCAB5* and *CCDC55*, as well as *SLCA6A4*. More distant genes, located within 250kb from the association peak include *TAOK1*, *GIT1*, *ANKRD13B* and *CORO6*, *BLMH*, *TMIGD1* and *CPD*. The region contains a GWAS risk locus, rs3110496, which has recently been associated with height (Lango et al. 2010). Another association peak, falling within the suggestive range, corresponds to the one observed in BART and is located between *SLC22A23* and *C6orf145*.



**Figure 23** – Whole genome association plot for EBER-1 transcript. Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10} P$ -values displayed on the Y axis.



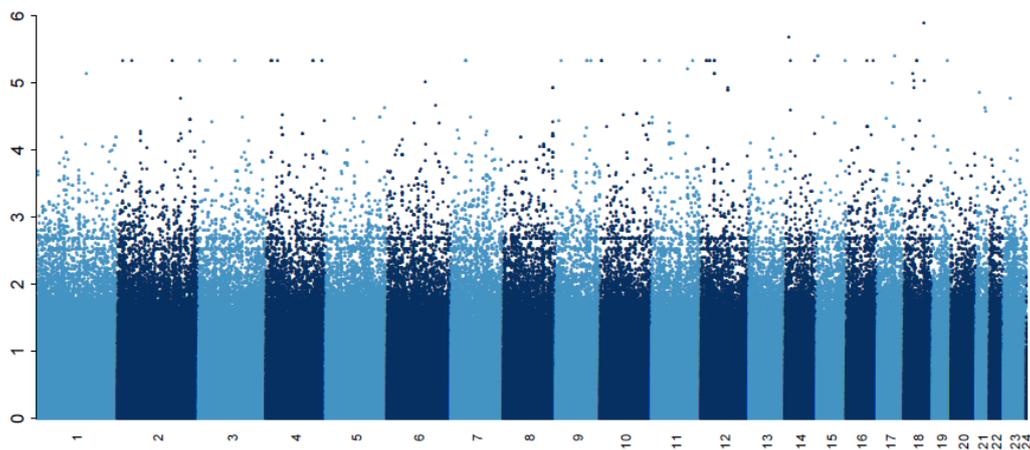
**Figure 24** – Regional SNAP plot for SSH2 locus showing association with EBER-1 expression. Genomic location is shown by the X-axis and P-value displayed on the Y-axis. Tested genotyped HapMap Phase 3 SNPs are depicted as squares. Most significantly associated SNP is shown by the red diamond and SNPs in LD are coloured in shades of red according to the extent of LD given by r-squared in to top left corner. Local genes are displayed at the bottom and their locations indicated by green lines.



**Figure 25** - EBER-1 expression by rs3809790 genotype – boxplot. Delta-Ct values displayed on Y axis. Values below zero indicate transcript abundance levels higher than in the housekeeper gene used to normalise the data.

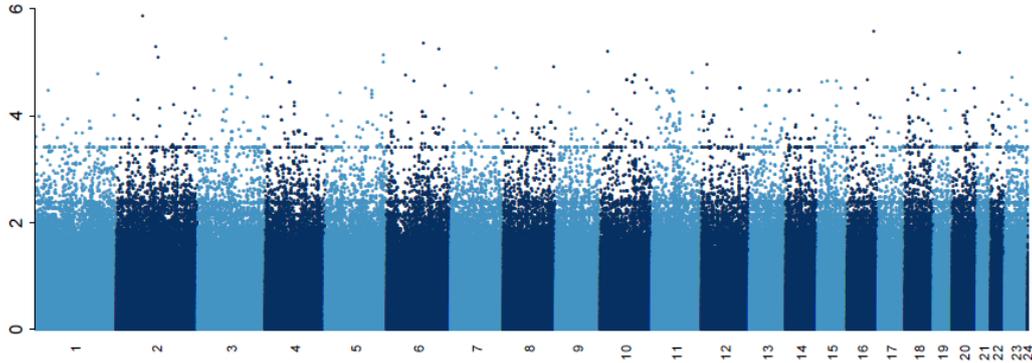
### 3.4.3. EBER2 and EBNA1

EBNA1 is a nuclear phosphoprotein which binds viral DNA circularising it into an episome which is required for the replication and maintenance of EBV genome (Young and Rickinson 2004 Kieff et al. 2010). These transcripts failed to provide any evidence for genome-wide significant eQTLs (Figure 26).



**Figure 26** – Whole genome association plot for EBER-2 transcript. Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10}$  P-values displayed on the Y axis.

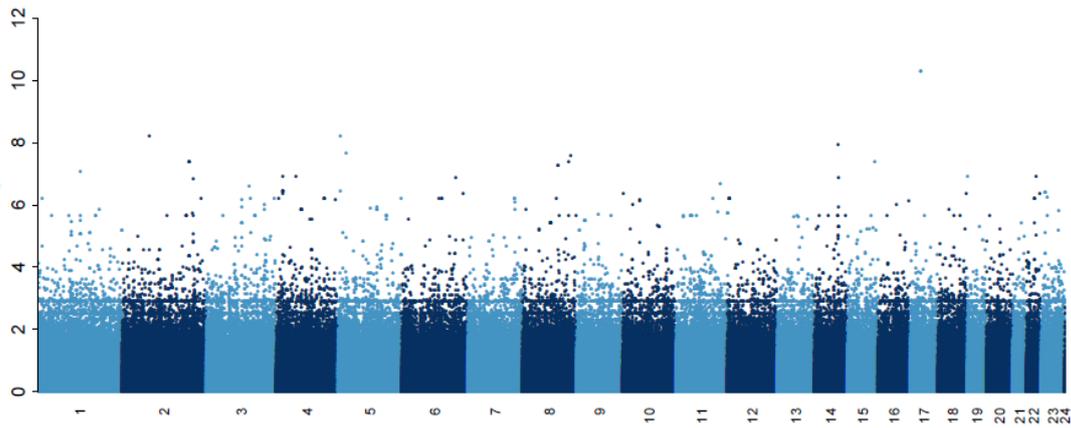
The only significant association for EBNA-1 was rs1453846 on chromosome 15 (P-value:  $3.61E^{-08}$ ) (Figure 27).



**Figure 27** – Whole genome association plot for EBNA-1 transcript. Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10}$  P-values displayed on the Y axis.

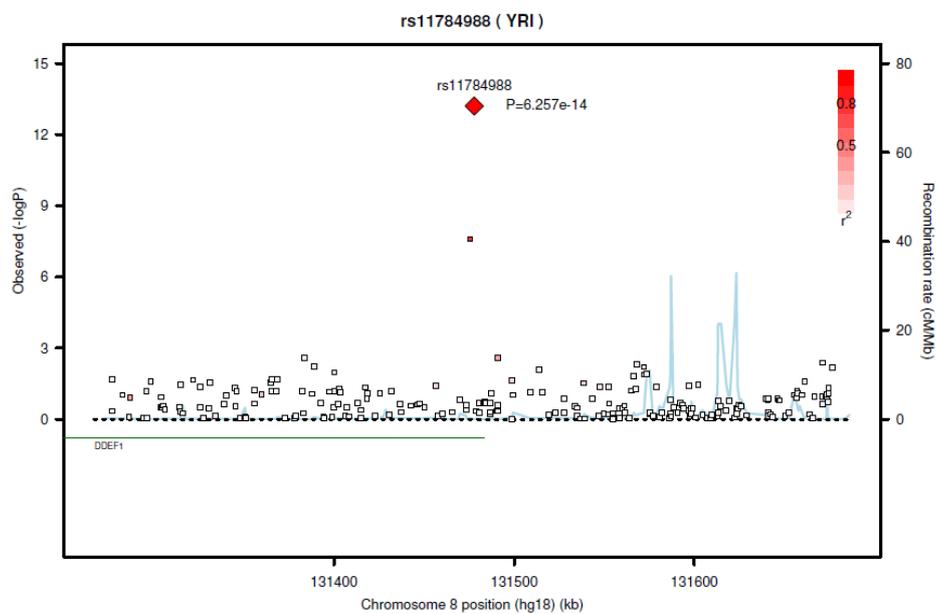
#### 3.4.4. EBNA2

EBNA2 is a phosphoprotein and the main EBV trans-activator for both cellular and viral genes (Young and Rickinson 2004 Kieff et al. 2010). By transactivating the viral Cp promoter it switches full latency gene expression observed early in B-cell infection (latency III) (Kieff et al. 2010). Suggestive associations were found located on chromosome 2, within a gene desert with no genes within approximately 500 kb on either side (Figure 28-30). Significant associations (rs11784988 and rs11785806, P-value  $6.26E^{-14}$ ) were found located on chromosome 8 within *ASAPI*, harbouring an MS risk locus (Bahlo et al. 2009), which however has not been replicated (Sawcer et al. 2011). There were other suggestive associations on chromosome 8 within *PLEKHF2* and close to radiation response QTLs (Niu et al. 2010) and to a type 2 diabetes risk locus (Voight et al. 2010). Weaker associations point to *ACCN1* locus on chromosome 17.

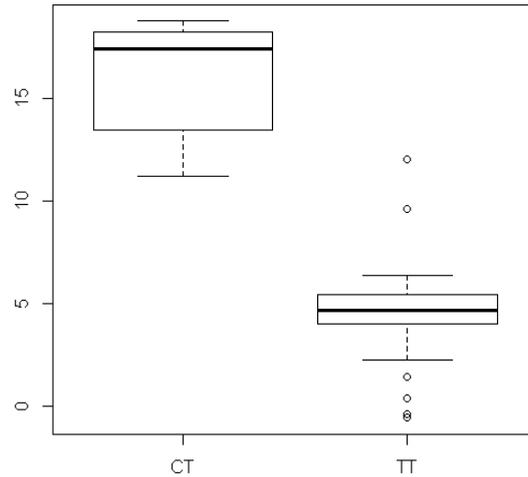


**Figure 28** – Whole genome association plot for EBNA-2 transcript. Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10} P$ -values displayed on the Y axis.

The association for ASAP1 consists of three SNPs (including two in high LD, Figure 29), however it is accompanied by a differentiation in levels of expression of EBNA-2 (Figure 30).



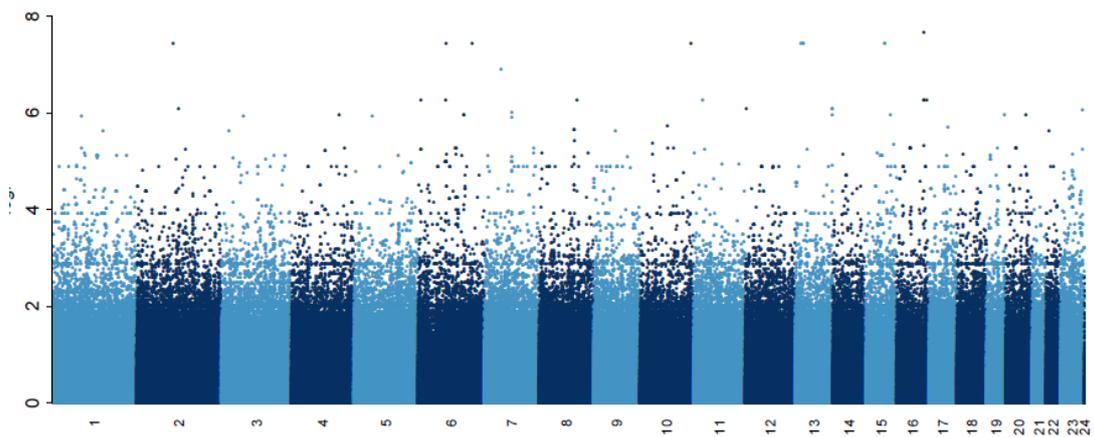
**Figure 29** – Regional SNAP plot for EBNA-2 associated region centered on rs11784988.



**Figure 30** – EBNA-2 expression by rs11784988 genotype – boxplot. Delta-Ct Values displayed on Y axis.

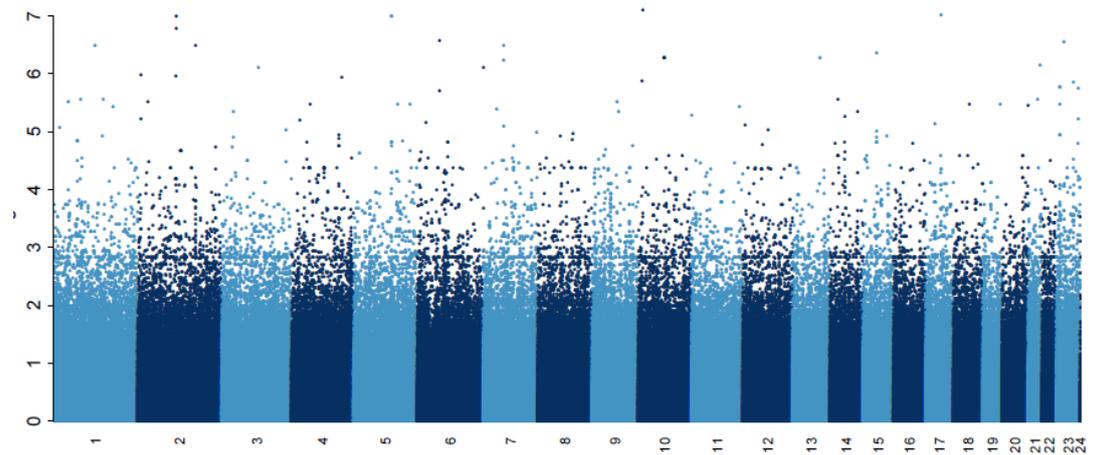
### 3.4.5. EBNA3 family

EBNA3 proteins are transcriptional regulators which control EBNA2 activity by repressing EBNA2-mediated transactivation (Young and Rickinson 2004). EBNA2 and the EBNA3 cooperatively control RBP-Jkappa activity regulating the expression of cellular and viral promoters (Young and Rickinson 2004). Associations were found localising to *DCLK1* on chromosome 13 (Figure 31), a locus containing SNPs associated with height, vertical cup-disc ratio and heart rate variability (Ndiaye et al. 2010, Ramdas et al. 2010, Newton-Chech et al. 2007). Also another putative eQTL associated with *THSD4* on chromosome 15, a gene containing a marker implicated in pulmonary function (Repapi et al. 2010). Two other smaller peaks correspond to the two previously observed in BART.



**Figure 31** – Whole genome association plot for EBNA-3A transcript. Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10}$  P-values displayed on the Y axis.

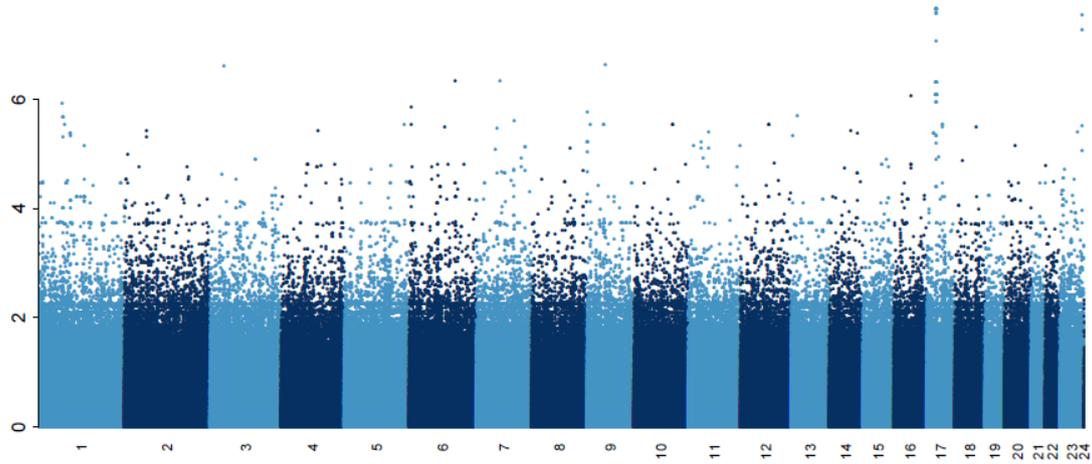
Expression of EBNA3B was associated with SNPs on chromosome 5 (Figure 32) within an intergenic fragment containing SNPs implicated in Sudden Cardiac Arrest (Aouizerat et al. 2011 Bhatnagar et al. 2011).



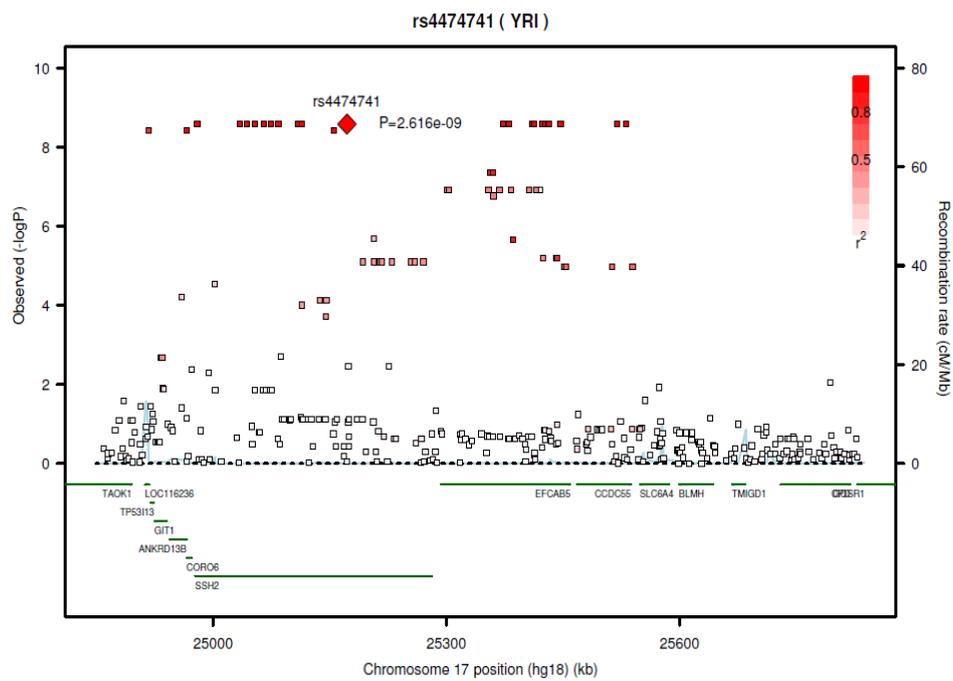
**Figure 32** – Whole genome association plot for EBNA-3B transcript. Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10}$  P-values displayed on the Y axis.

The associated region for EBNA3C encompasses *SSH2* and the neighbouring *EFCAB5* on chromosome 17 (P-value 2.22E-08) (Figure 33-35). Another SNP associated with EBNA3C

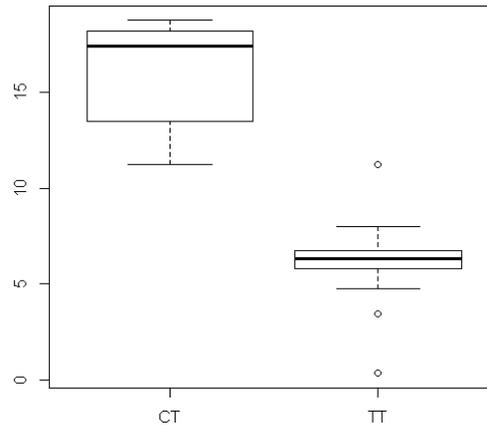
is located on chromosome X. Within 200kb of the SNP there is *SPANXN1* as well as rs6627057, associated with bipolar disorder and schizophrenia (Wang et al. 2010).



**Figure 33** – Whole genome association plot for EBNA-3C transcript. Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10} P$ -values displayed on the Y axis



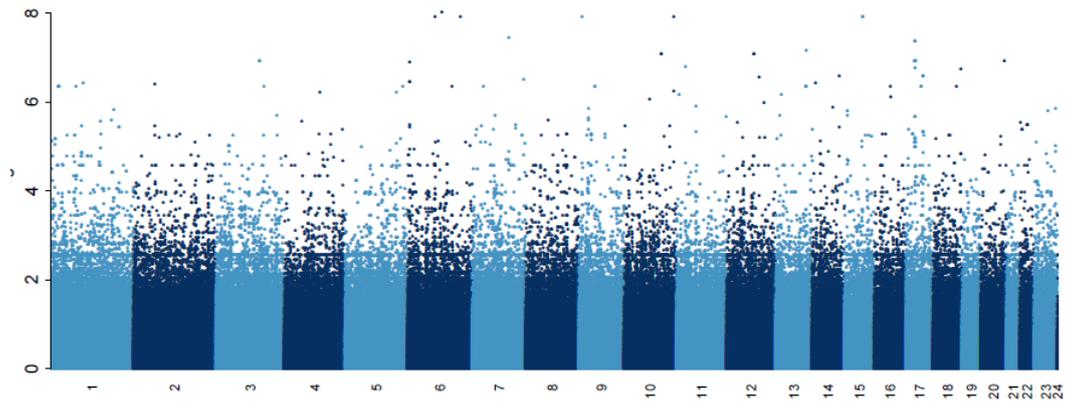
**Figure 34** – Regional SNAP plot for SSH2 locus and association with EBNA3C expression.



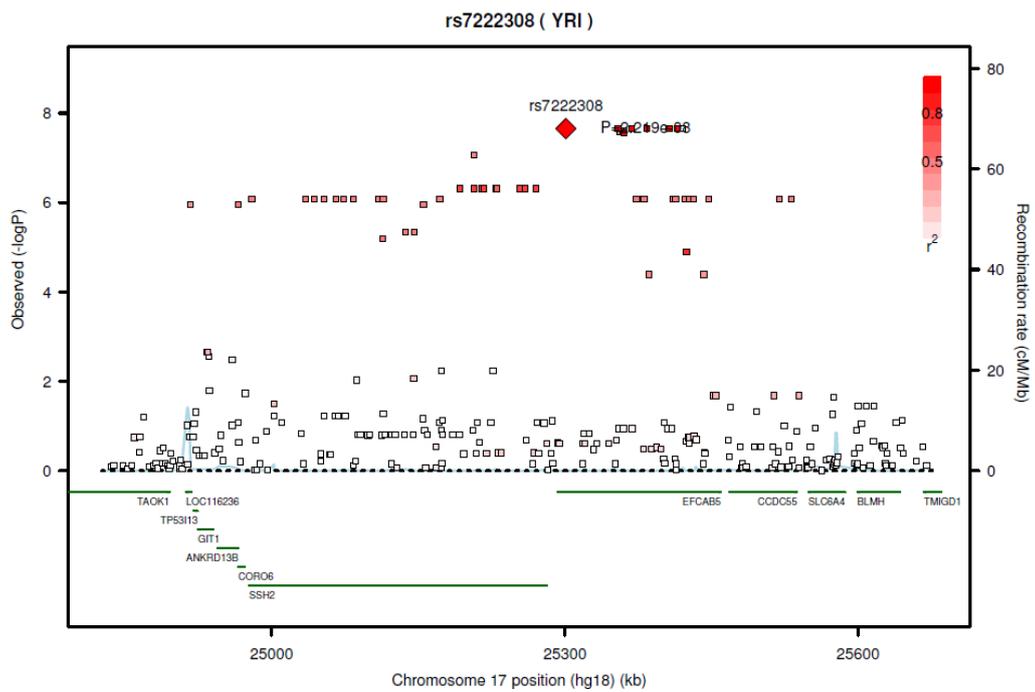
**Figure 35** – EBNA-3C expression by rs4474741 genotype- boxplot. Delta-Ct Values displayed on Y axis.

### 33.4.6 LMP1

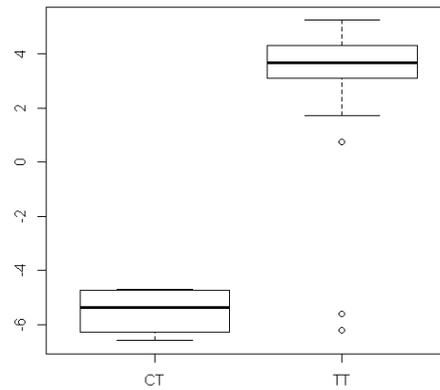
LMP1 resembles CD40 and can partially substitute for it *in vivo* providing both survival and differentiation signals necessary for the B-cells to progress through the germinal centre and differentiate into resting memory B-cells. Associations found include SNPs localising to *SLIT1* on chromosome 10 and *SSH2* on chromosome 17 (Figure 36-38). Other associations include the region of *THSD* (Chr 15) as well as an intergenic region on chromosome 12. There was a number of solitary associations with one significant SNP encompass *SLIT1*, *AK3*, *PIK3CG*, *SLC1A1* and *FBXO10 / TOMM5* and *NALCN*. The hypoxia study LMP1 eQTL (rs1913243 ) has not been available since it has not passed the QC, however no significant association were present at the locus.



**Figure 36** – Whole genome association plot for LMP-1 transcript. Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10}$  P-values displayed on the Y axis.



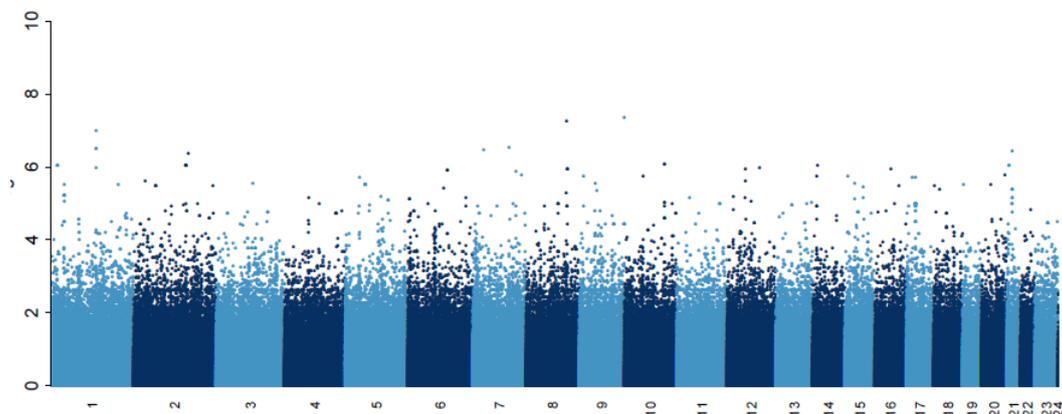
**Figure 37** – Regional SNAP plot for SSH2 locus and association with LMP1 expression.



**Figure 38** – LMP-1 expression by rs7222308 genotype – boxplot. Delta-Ct Values displayed on Y axis.

### 3.4.7 LMP2

LMP2A is a transmembrane protein that contains immunoreceptor tyrosine-based activation motifs and thus mimics the actions of BCR receptor, delivering a non-proliferative signal necessary for B-cell maturation, resembling that of an intact BCR (Thorley-Lawson 2001 Young and Rickinson 2004). The analysis yielded no significant association peaks for LMP2. Four statistically significant single-SNP associations were located in different intergenic regions with no evident function (Figure 39).



**Figure 39** – Whole genome association plot for LMP-2 transcript. Chromosomes are displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10}$  P-values displayed on the Y axis.

### **3.4.8 EBNA-LP**

Despite testing over a range of annealing temperatures, the primers for EBNA-LP failed to yield a specific product described by Pan et al. (2005). Owing to the highly repetitive structure of the transcript and its similarity to EBNA-2, successful amplification can be challenging (personal correspondence with Dr Paul Ling, former officer at the International Association for Research on EBV and Associated Diseases). For the reasons described, EBNA-LP had to be excluded from the current and subsequent analysis.

### **3.5 EBV latency eQTLs in the MRC-A panel**

In order to improve the power of the association tests, the search for latency eQTL was extended to a larger cohort MRC-A from the asthma eQTL study (Moffat et al. 2007). This cohort consisted of 57 singletons, 109 sib pairs, 23 sib trios and 2 quadruplets from 180 British families (352 LCLs in total before cDNA conversion and QC, see Appendix). To account for the family-based character of the tested cohort, a statistical tool Merlin (accepting an externally provided kinship matrix as input) was used which can implement family-based association tests (Chen 2007). More specifically, merlin—offline function was applied because of its ability to work with imputed genotype data. This was necessary because HapMap Phase 3 MACH-imputed genotypes were used for association tests.

Genotypes for the MRC-A panel were obtained for the Illumina Sentrix Human-1 Genotyping BeadChip (100k), Illumina Sentrix HumanHap300 BeadChip (300k) as well as in the HapMap Phase 3 MACH-imputed format. The transcripts were quantified using SybrGreen qPCR method and the same primers as previously (section 3.4). As a new cohort was used for the assay, and, unlike in the hypoxia study panel, all MRC-A LCLs represented

the 1<sup>st</sup> passage of growth, it was decided to test a selection of housekeeping genes for consistency of expression and choose the best gene(s) for results normalisation.

### 3.5.1 Housekeeper assay

Housekeeping genes are considered to have stable expression levels across different cells of the same type and, in some instances, also across different cell lines and thus are commonly used as a yardstick by which the expression of the gene of interest (which may vary from cell line to cell line) is measured (Vandesompele 2002, Hugett 2005, deJonge 2007). This allows for comparison of the data from different reactions/individuals irrespective of the differences in the input RNA used in these reactions (Pfaffl et al. 2004, de Brouwer et al. 2006). However, housekeeping gene stability may vary and it is essential to determine which genes constitute the best candidates for a particular set of samples and experiment (Radonic et al. 2004, Dheda et al. 2004, Vandesompele 2002, Pfaffl et al. 2004). Since *OAZI* had been previously identified as the most stably expressed housekeeping gene in the hypoxia study cohort (Mohr 2010), therefore it was retained for the EBV latency assays conducted using the hypoxia study samples. However, for the new MRC-A cohort another optimisation of housekeeping genes was necessary.

Six genes, *GAPDH*, *BACT*, *OAZI*, *HPRT1*, *RPL60*, *RPS6*, commonly used in RT-PCR and described in the literature (de Brouwer et al. 2006, Lord et al. 2010, de Jonge et al. 2007), were tested for stability across a random set of 50 HapMap LCLs. Two algorithms were used, Bestkeeper and SLqPCR (a version of geNorm written in R) which are independent statistical applications. The best pairs with highest pairwise correlation coefficient and lowest variation were *GAPDH* + *HPRT1* (selected with SLqPCR) and *RPL60* + *RPS6* (selected using Bestkeeper with correlation coefficient of over 0.9). However

Bestkeeper, which takes into account also the sample-to-sample variability in Ct values, indicated that *GAPDH* and *HPRT* have much higher SDs and therefore are less suitable. This was also evident from the raw data. Therefore it was decided that the two ribosomal housekeepers, which also appeared as second-best pair in geNorm, would be retained as the best combination of stable reference genes for normalising expression of EBV transcripts (Mohr 2010).

### **3.5.2 Latency eQTL assay results**

RNA from 352 MRC-A LCLs was reverse transcribed into cDNA and assayed for expression levels of 10 latency transcripts. Primers for the 11<sup>th</sup> transcript ENBA-LP failed the specificity test (as stated before).

#### **Quality Control**

MDS analysis was performed on Illumina 300k MRCA genotypes to check for possible population outliers (according to MDS protocol given in the Appendix). The samples' genotypes were tested together with corresponding HapMap Phase 3 Yoruban CEPH and Chinese genotypes used as reference. No individuals were excluded as population outliers.

Reactions were conducted in duplicates and, if the SD of the replicates exceeded 1.5, repeated again in triplicates (up to two more times). Samples whose SD exceeded 1.5 repeatedly were removed from further analysis. Two housekeeping genes *RPL30* and *RPS6* were used for data normalization (in order to obtain delta-Ct values equivalent to log-transformation). The two reference genes revealed almost identical expression levels with very low SD (below 0.4) whose magnitude was the same as between technical replicates of

each of the housekeeping genes on its own. However, in several reactions *RPS6* assay repeatedly yielded no detectable expression and therefore only *RPL30* was retained for the normalisation of Ct values. Since the two genes scored high in tests of stability by two algorithms, and their transcript expression levels were highly similar (SD below 0.4), this should not have a significant impact on the quality of expression data.

Delta-Ct values were thus calculated using *RPL30* reference replicates only. Delta-Ct values lower or higher than two times standard deviation from the mean of the whole cohort were considered as outliers and removed from the analysis. This meant that before the tests were implemented, between 5 to 15 outliers were removed for each phenotype tested. An alternative approach would be to apply a conservative quantile normalisation to enforce normal distribution. The delta-Ct values represent log-transformed phenotypes and no other transformation was applied before the first round of tests. In the end however, Z-score normalisation was applied to the tested EBV latency phenotypes. The reason for this was that non-normalised delta-Ct phenotypes were still non-normally distributed even after removing the most obvious outliers (two times standard deviation from the mean), and this resulted in numerous solitary associations (Table A4, Figure A5), particularly abundant for EBNA3B. Quantile normalization was also considered, however it would likely leave no eQTL candidates because of its high stringency. Filters included 10% for individual and SNP missing rate, P-val HWE<0.001 and MAF > 0.01.

PCA of the EBV expression data was conducted and the test was repeated multiple times with a variable number of 0-5 PCs adopted as covariates to control for possible batch effects, and population structure or cryptic relatedness (merlin—offline takes pre-specified familial relatedness into account when calculating kinship matrix). No PC-covariates correlated with RNA/cDNA conversion plate, qPCR plate or sex and family ID, however since a possible

batch effect was observed when correlating Ct values to the original boxes in which the RNA was first shipped and provided from Professor Cookson lab and using 1-3PCs did not alter the main association results, and using 3PCs seemed to result in most significant associations, therefore data obtained with 3 PC-covariates was analysed and presented in Table 9 (but not in Table A4 where delta-Ct phenotype results with no PC correction are shown). HapMap3 imputed genotypes were not available for 56 samples and 19 samples yielded no expression, which reduced the number of LCLs available for the association test to 277, however some additional samples (5 to 15) had to be removed in each test because their corresponding phenotypes differed by more than 2 SDs from the mean.

No QC has been conducted on the MRCA genomewide expression data provided by Professor William Cookson except for an additional PCA (see section 5.5.3).

## **Results**

The most significant associations are summarised in Table 9.

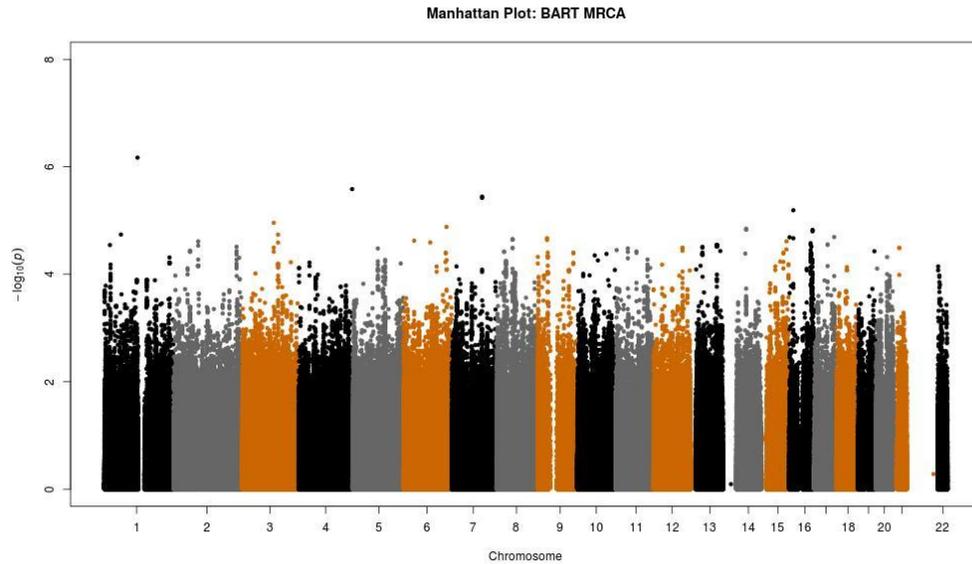
<i>SNP</i>	<i>Chr</i>	<i>Solitary</i>	<i>MAF</i>	<i>TRAIT</i>	<i>P-value</i>	<i>Gene</i>	<i>eQTL / disease association</i>
rs17158616	10		+	EBNA1	1.06E-08	RASGEF1A	
rs17158630	10		+	EBNA1	1.10E-08	RASGEF1A	eQTL for CD83 - interacts with EBV's LMP1
rs17339199	7	+	+	EBNA3A	2.12E-08	FKBP6	eQTL for IL12B - MS and lymphoma associated
rs11867159	16	+	+	EBER1	6.73E-08	RBFOX1	
rs12660137	6			EBER1	6.77E-08		eQTL for TNFRSF1B - MS SLE PTLD associated

**Table 9** – EBV latency eQTL candidates. Only the most significant associations at  $P\text{-value} < 1E^{-07}$  are listed

- i) SNP – statistically significant associations are listed by SNP in the first column.
- ii) Chr – gives the genomic location of a significant association by chromosome
- iii) Solitary – “+” sign signifies that no other SNP associated at  $p < 1E^{-07}$  was present within 250kb
- iv) MAF – “+” denotes whether minor allele frequency of the associated SNP exceeded 0.5
- v) TRAIT – lists the EBV latency transcript associated with a relevant SNP
- vi) P-value – reports the P-value of the test statistic
- vii) Gene – reports the genomic location of the associated SNP by gene
- viii) eQTL – lists human genes whose expression had been associated with a given SNP in previous eQTL mapping experiments included in the SCAN, Genevar and GTEx databases.

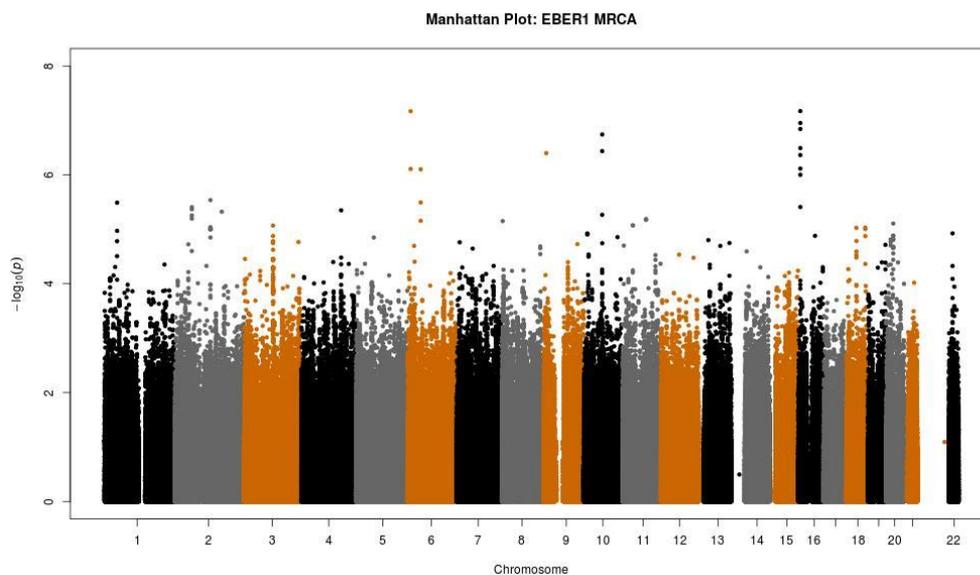
The associations were prioritised on the basis of statistical significance; MAF; number of significantly/suggestively associated SNPs within a single 400kb window centered on the SNP with lowest P-value; genomic location of the candidate eQTL and its function (if within or near a functional element); overlap between the candidate viral eQTL and human eQTLs obtained from the same MRC-A cohort by a global gene expression assay; overlap between the candidate viral eQTL and human eQTLs sourced from public databases (SCAN, Genevar and GTEx); and functions and associations of human genes reportedly affected by the candidate viral eQTL. Only three SNPs reached the nominal genome-wide significance level, and two of them were located in the same locus (within the ORF of *RASGEF1A*) and appeared as solitary associations on the regional association plot. The SCAN database identified three of the most significantly associated SNPs as eQTLs for human transcripts. SCAN's database includes eQTLs from 87 HapMap CEU and 89 YRI LCLs assayed with Affymetrix Human Exon 1.0 ST Array (targeting 60 000 transcripts) (<http://www.scandb.org/>). No overlap was observed with significant eQTLs from the Genevar database sourced from an experiment conducted by Stranger et al. (2012) on 726 HapMap LCLs from 8 different human populations (<http://www.sanger.ac.uk/>). Interestingly, all three human transcripts had a biological link either to EBV infection or EBV-related autoimmune diseases. No overlap with MRCA eQTLs was present.

No significant associations were present for BART small untranslated RNAs (Figure 40) and the overall association was low and below  $1 \times 10^{-6}$ .

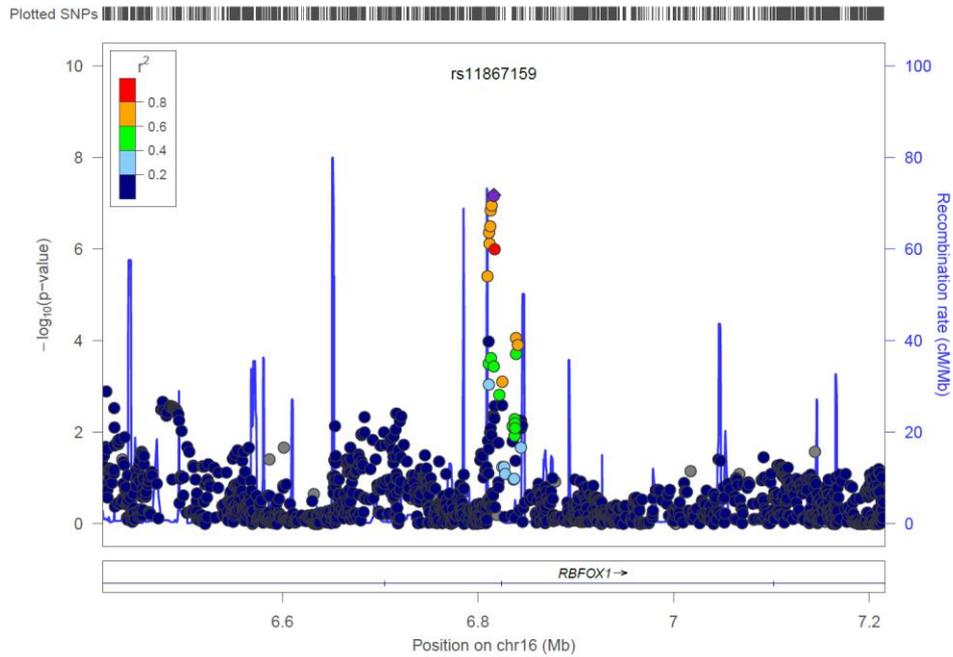


**Figure 40** – BART genomewide associations. Chromosomes are displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis.

For the EBER1 untranslated RNA, a single association peak with a significant eQTL (rs11867159,  $6.7E^{-08}$ ) was identified on chromosome 16 within the ORF of *RBFOX1* (Figure 41-42). Another solitary significant association was present on chromosome 6.

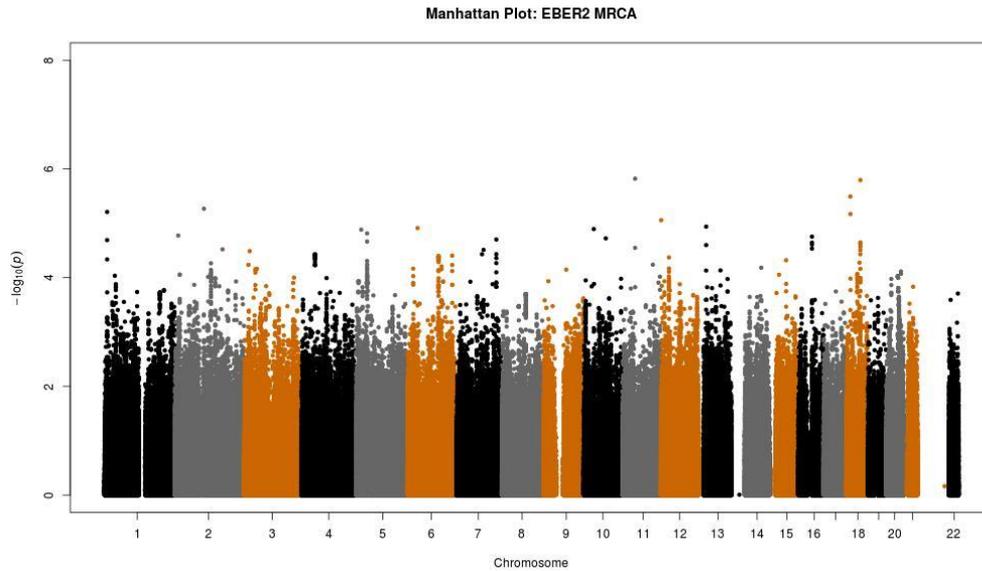


**Figure 41** - EBER1 genomewide associations. Chromosomes are displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis.

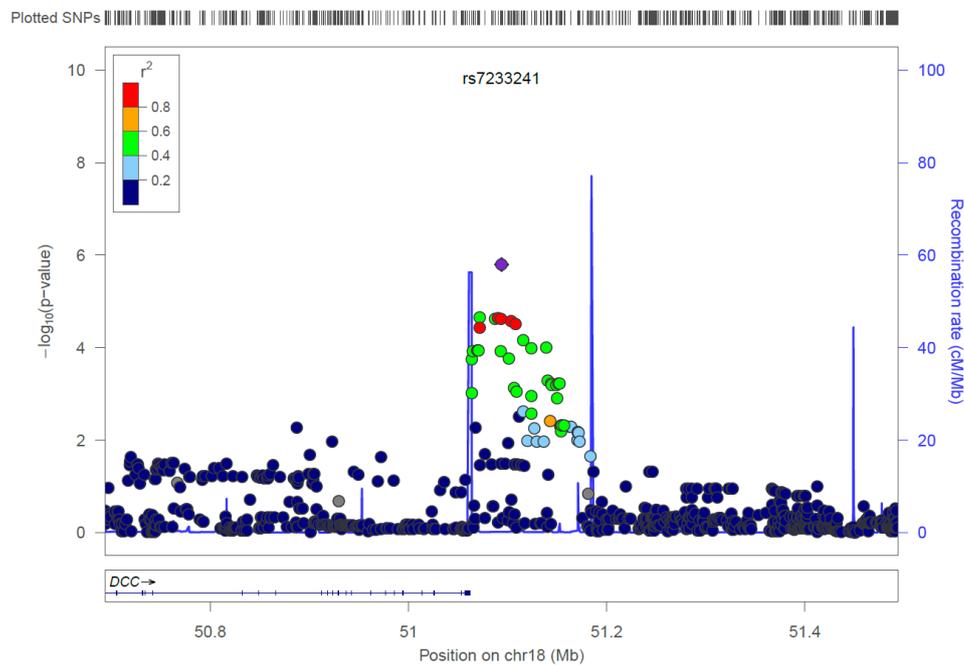


**Figure 42** – EBER 1 and rs11867159 (RBF0X1) regional association plot; rs11867159 indicated by a purple diamond; Genomic location indicated by the X-axis, P-value by the Y-axis;  $r^2$  scale showing the degree of LD between the lowest P-value SNP other SNPs displayed in the upper left hand corner; ; only associations at  $p < 1E^{-01}$  are shown.

No significant associations were found for the other EBV-encoded small RNA. A small peak with a suggestive association (rs7233241,  $1.60E-06$ ) was present on chromosome 18, proximal to an immunoglobulin family transmembrane receptor (Figure 43-44), Deleted in Colorectal cancer (*DCC*).

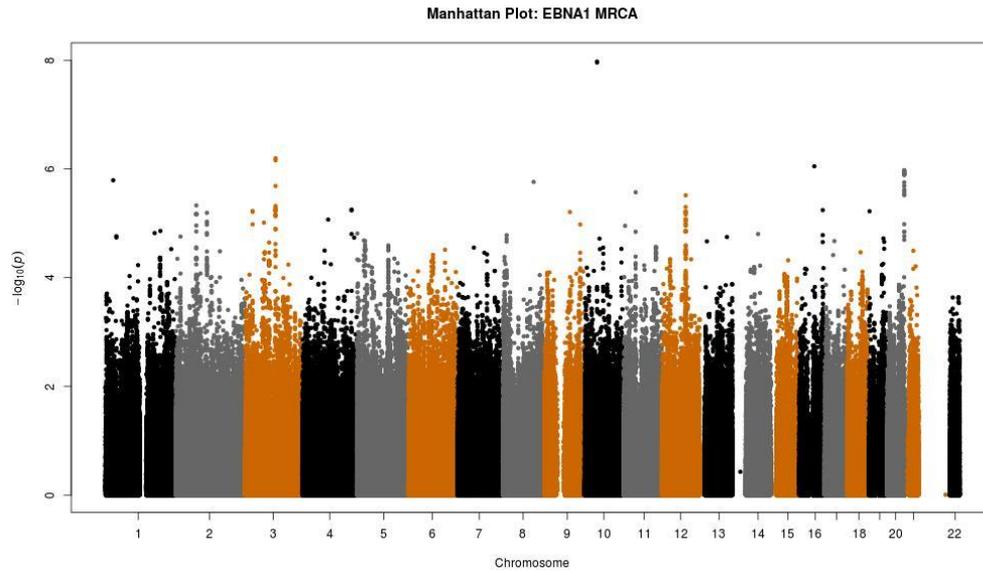


**Figure 43** - EBER2 genomewide associations. Chromosomes are displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis.

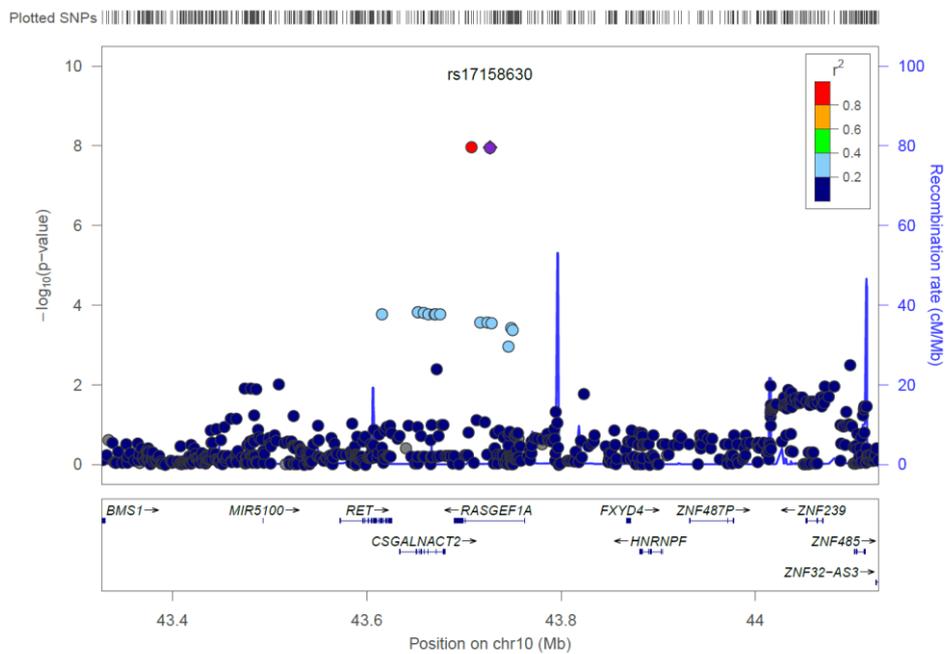


**Figure 44** – EBER2 and rs7233241 (DCC) association to EBE2 regional association plot; rs7233241 indicated by a purple diamond; only associations at  $p < 1E^{-01}$  are shown.

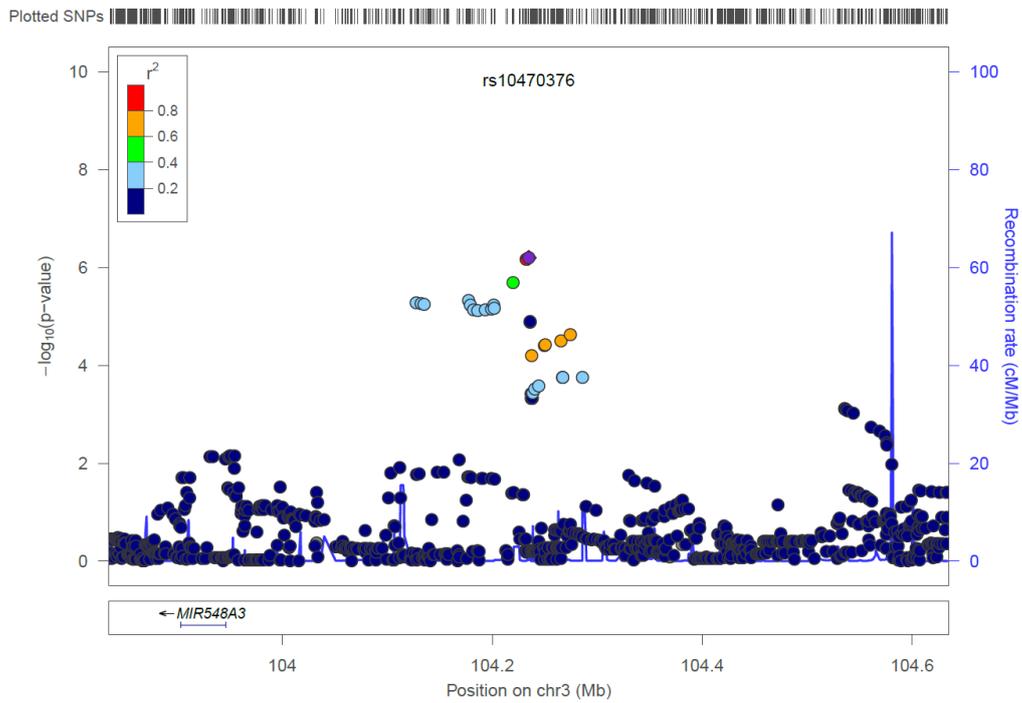
Two SNPs significantly associated with EBNA1 levels were located on chromosome 10, within the ORF of *RASGEF1A*, however did not constitute a part of a continuous association peak (Figure 45-46). Suggestive association peaks for the EBNA1 were also located on chromosome 3 and chromosome 20 (Figure 47-48).



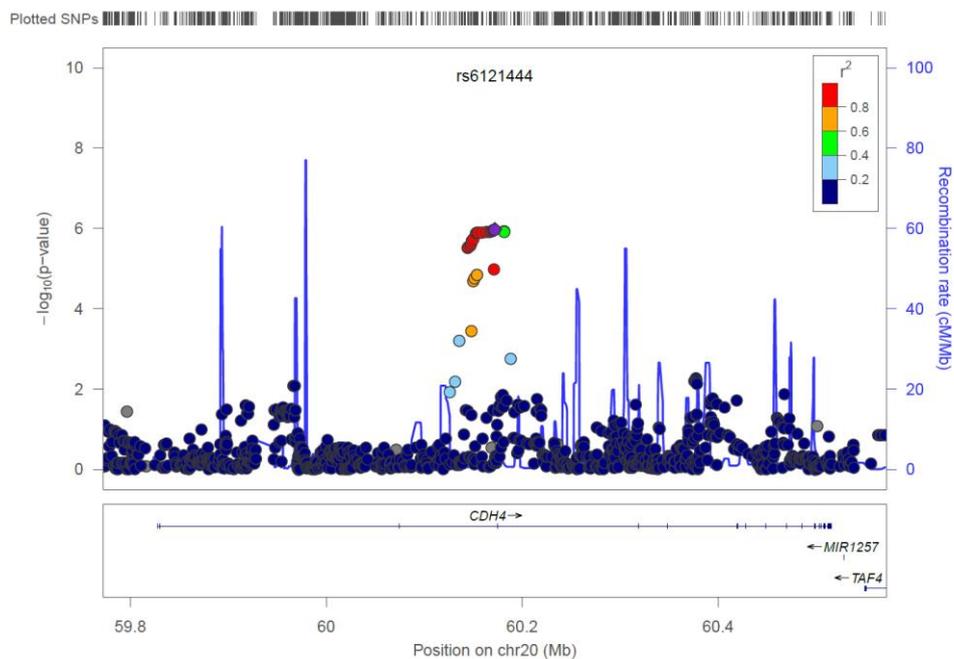
**Figure 45** - EBNA1 genomewide associations. Chromosomes are displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis.



**Figure 46** - RASGEF1A and rs17158630 association (EBNA1) regional plot; only associations at  $p < 1E^{-01}$  are shown.

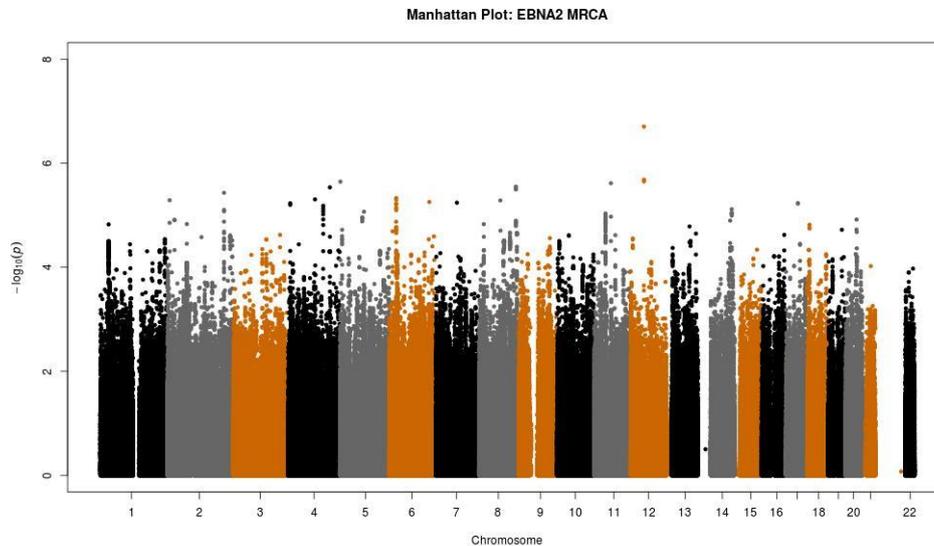


**Figure 47** – EBNA 1 and rs10470376 regional association plot. rs10470376 indicated by a purple diamond; only associations at  $p < 1E^{-01}$  are shown.



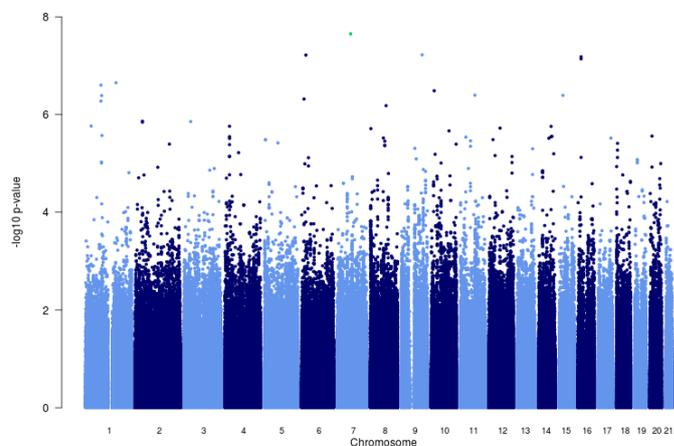
**Figure 48** - EBNA1 and rs6121444 (CDH4) regional association plot; rs6121444 indicated by a purple diamond; only associations at  $p < 1E^{-01}$  are shown.

The peak on chromosome 20 fell within *CDH4*, a member of the cadherin family of proteins regulating cell-cell adhesion which has recently been proposed to be a new putative tumor suppressor gene epigenetically silenced in NPC (Du et al. 2011). In contrast, weak associations were present for the EBNA2 (Figure 49).



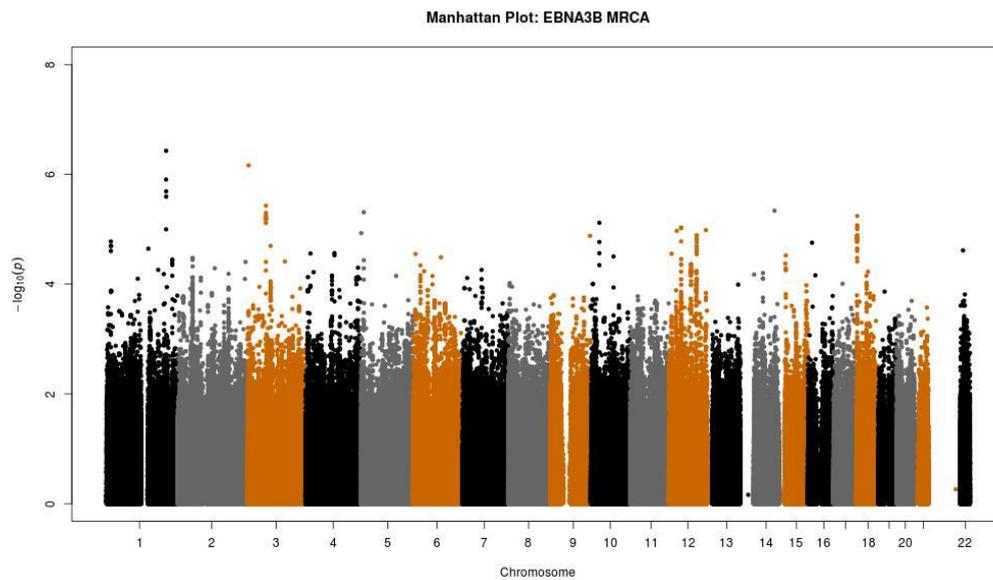
**Figure 49** - EBNA2 genomewide associations. Chromosomes are displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis.

A single significant association (rs17339199, within *FKBP6*) was identified by the EBNA3A assay (Figure 50). The SNP however, was not a part of an association peak.



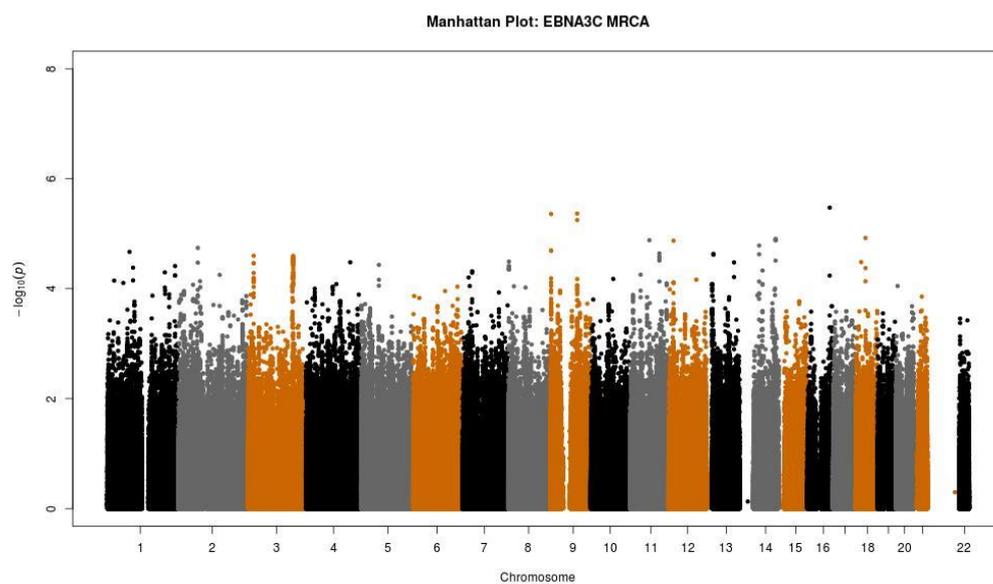
**Figure 50** - EBNA3A genomewide associations. Chromosomes displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis. Significant SNP marked in green

No significant associations for EBNA3B transcript levels were found (Figure 51).



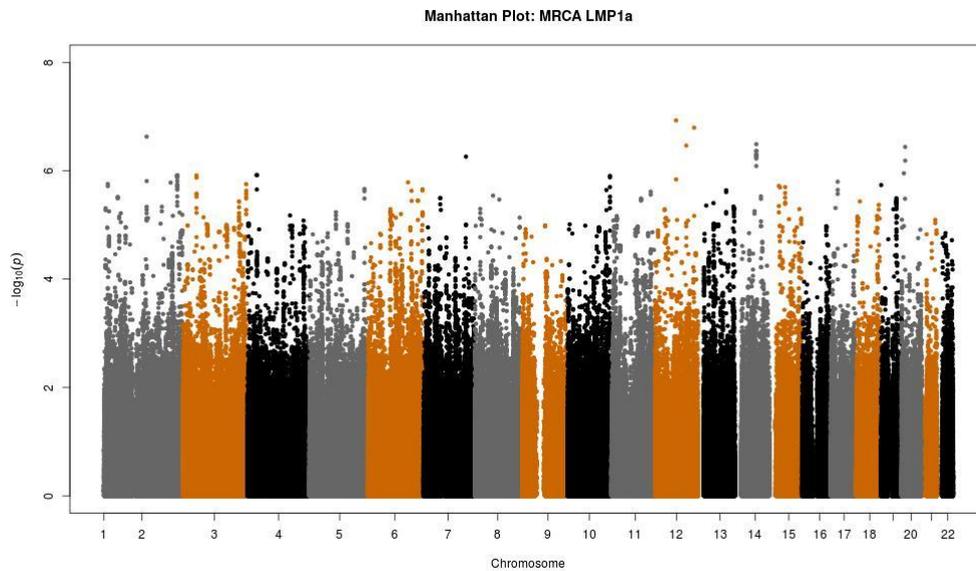
**Figure 51** - EBNA3B genomewide associations. Chromosomes are displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis.

No significant associations have been identified for the EBNA3C transcript (Figure 52).



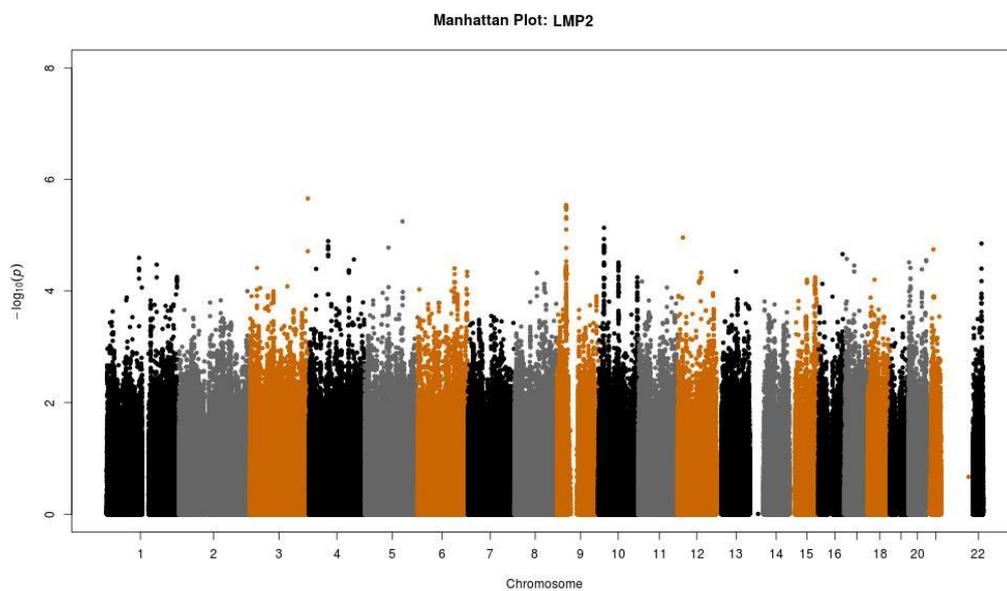
**Figure 52** - EBNA3C genomewide associations. Chromosomes are displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis.

The overall association levels of LMP1 were higher with more random noise present, than in the other latency transcripts, however no single association reached genomewide significance (Figure 53).



**Figure 53** - LMP1 genomewide associations. Chromosomes are displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis.

No significant association results were obtained for the LMP2 expression levels (Figure 54).



**Figure 54** - LMP2 genomewide associations. Chromosomes are displayed along the X axis;  $-\log_{10}$  P-values displayed on the Y axis

### **3.5.3 eQTLs for both EBV and human transcripts**

The original study of global gene expression in the MRC-A panel LCLs measured the expression levels of 54,675 transcripts representing 20,599 genes (Dixon et al. 2007). By courtesy of Professor Cookson, the expression data (measured with the Affymetrix HG-U133 Plus 2.0 chip) was made available for the purpose of the current study. The expression data had already been subjected to QC conducted by Professor Liang and Professor Cookson. First, all gene expression was normalized together using the robust multi-array average (RMA) package in order to remove technical and spurious variation and background noise (Irizarry et al. 2003, Bolstad et al. 2003). A second inverse normal transformation step was also applied to each trait to avoid any outliers and enforce normal distribution. Finally the expression data was subjected to PCA correction. Since all microarray probes per individual are subjected to identical experimental conditions, it is possible to summarise them through a principal components analysis which regresses out noise in the form of the top principal components of gene expression that correlated with technical factors like RNA extraction and cDNA synthesis dates, the date that the sample was fragmented, and the date of chip hybridization. Human expression data used for this experiment has been subjected to a new method for dimension reduction which accounts for nongenetic effects in estimates of expression levels, developed by Liang et al. (2013). Like the Bayesian factor analysis model used by Pickrell et al. (2010) this method models all the unobserved confounders explicitly (Liang et al. 2013). And likewise, this method selects such number of PCs that that results largest number of eQTLs (Liang et al. 2013). The expression data was then made available for the current study.

Datafiles with array feature locations and full array design annotation were downloaded using accession number (E-MTAB-1425) for the MRC-A panel from the ArrayExpress online directory at European Molecular Biology Laboratory - European Bioinformatics Institute,

EMBL-EBI website and from the Affymetrix website. Processed expression data was then tested against HapMap 3 imputed genotypes multiple times, each time up to 50 PCs used as covariates to regress out any potential residual confounding effects. As predicted, using 0 to 3 PCs as covariates yielded most associations. This is because PCA correction had already been conducted and applied previously (as discussed above). Consequently results obtained with 0 PC-covariates were used for the purpose of this study. Results are summarised in Table 10. The results of the assay were contrasted with the list of candidate EBV eQTLs. No SNP-probe association turned out to cross the nominal genome-wide significance threshold although a single suggestive association was present to an uncharacterised transcribed locus (shown in bold in Table 10).

SNP	EBV transcript	P-value	Locus	Affymetrix probe	P-value	human transcript
rs17158616	EBNA1	1.06E-08	RASGEF1A			
rs17158630	EBNA1	1.10E-08	RASGEF1A			
<b>rs17339199</b>	<b>EBNA3A</b>	<b>2.12E-08</b>	<b>FKBP6</b>	<b>1565565_at</b>	<b>5.83E-07</b>	<b>Human transcribed locus. Represented by 2 ESTs from 1 cDNA libraries.</b>
rs11867159	EBER1	6.73E-08	A2BP1			
rs12660137	EBER1	6.77E-08				

**Table 10** – Human MRC-A eQTLs. Significant EBV eQTL candidates (SNPs) and their respective EBV transcripts are listed in the first two columns. The last three columns list the associated Affymetrix probe, P-value for the test statistic and the corresponding human transcript linked to the same candidate EBV eQTL by the association test.

### 3.6 EBV eQTLs from RNA-seq experiments

In 2013, a study by Arvey et al. (2013) investigated EBV expression and interactions mining RNA-seq expression data from previous experiments by Birney et al., 2007; Cheung et al., 2010; Kasowski et al., 2010; Montgomery et al., 2010; and Pickrell et al., 2010. The group

mapped transcript sequences to the EBV genome, quantified the expression of latency and lytic genes, and investigated correlations between human and viral expression identifying groups of genes that are co-expressed. Associations between EBV transcript levels and human genetic variation were beyond the scope of study and not discussed, however, EBV expression data combining, 83 latency and lytic transcripts, was made public (<http://ebv.wistar.upenn.edu/download.html>) and available for further investigation. That data became the subject of the next analysis in the current search for potential EBV eQTLs.

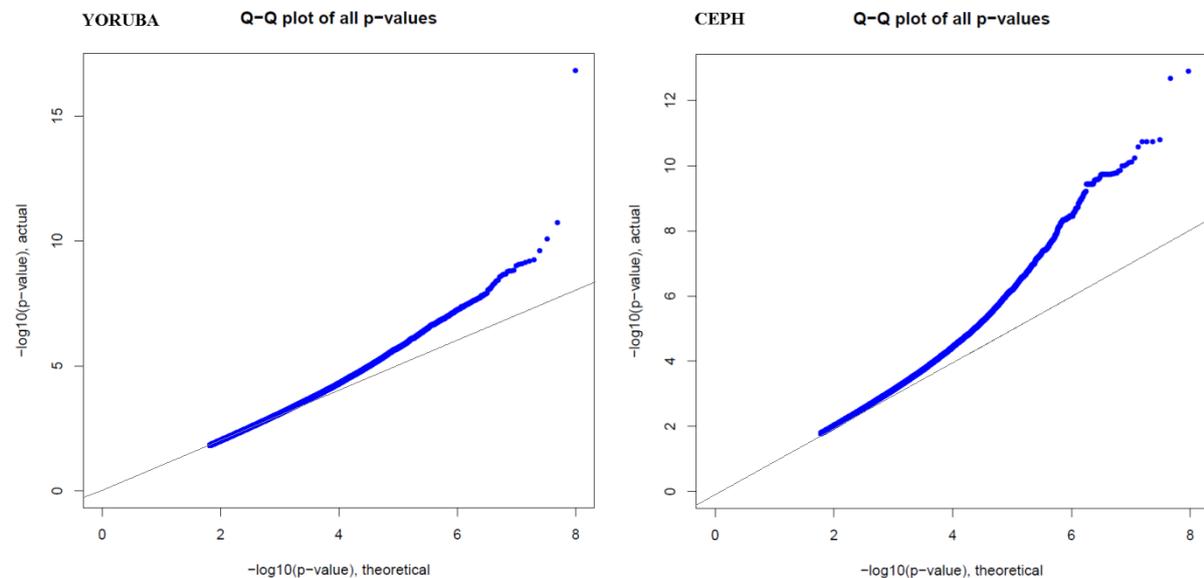
### **Quality Control**

EBV expression for 67 individuals from the HapMap Yoruban panel was sourced from the Pickrell et al. (2010) study (the viral expression was quantified from raw read by Arvey et al. 2011 and uploaded to <http://ebv.wistar.upenn.edu/download.html>), while viral expression data for 67 CEPH individuals was pooled from non-overlapping LCLs grown by Cheung et al. (2010) and Montgomery et al. (2010). No further phenotype transformation was performed. Genomic information was obtained from the HapMap Phase 3 download directory at the HapMap website (28-May-2010) and data for both populations was analysed separately using MatrixEQTL (with a MAF threshold of 5%, SNP missing rate 10%, P-val HWE<0.001) with PCA correction (between 0 to 5 PCs was used as covariates in the association test). The results discussed in the section below were obtained using no PCs.

#### **3.6.1 RNA-seq EBV latency and lytic eQTLs**

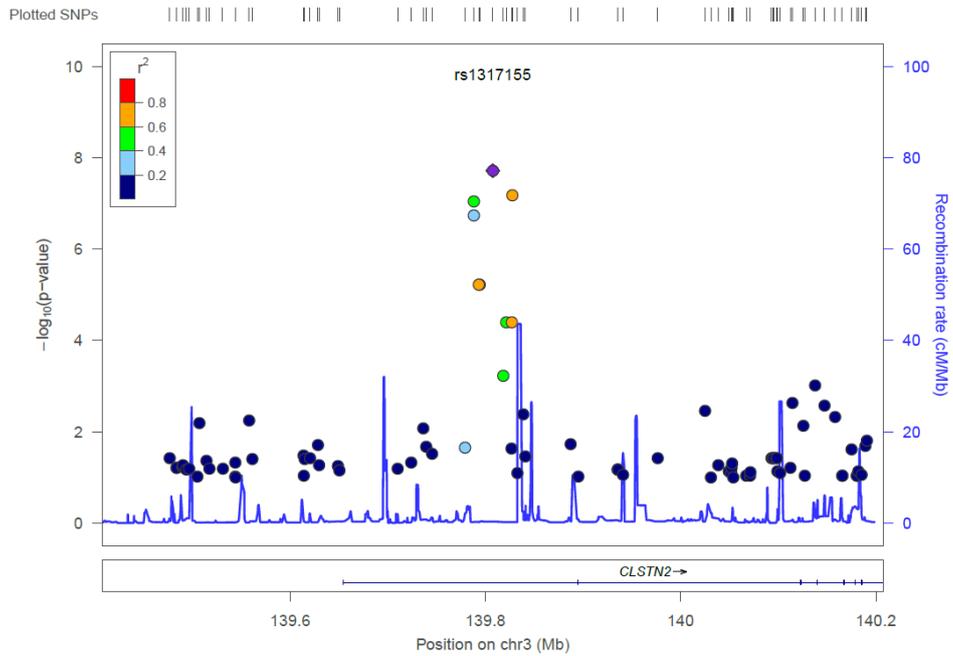
Initial analysis yielded a set of significantly associated SNPs which were prioritised depending on the strength of the association, number of significantly or highly (P-value of

$1 \times 10^{-6}$  and higher) associated SNPs in any 400kb region, their location, and whether they have already been identified as eQTLs in eQTL databases (SCAN, Genevar, GTEEx). QQ plots are shown for the YRI and CEPH cohorts (Figure 55).

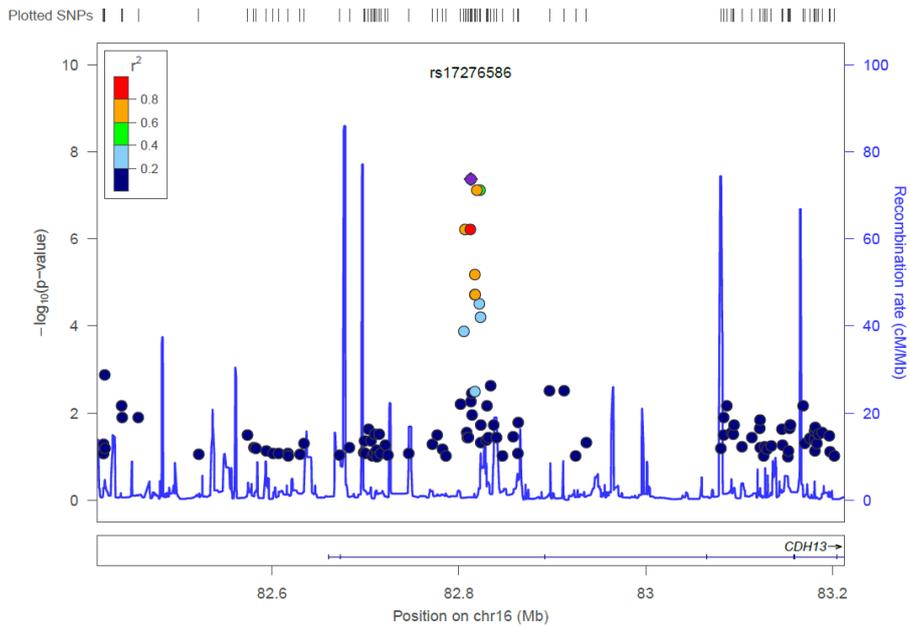


**Figure 55.** Genome wide QQ plots for RNA-seq EBV eQTL associations.

For YRI, the strongest eQTLs were found for EBNA1 involving three different genomic regions. These included rs1317155 (P-value  $1.9E^{-08}$ ) in the region of *CLSTN2* on chromosome 3 (Figure 56); rs17276586 (P-value  $4.2E^{-08}$ ) in region of *CDH13* on chromosome 16 (Figure 57) and rs17564816 (P-value  $5.2E^{-09}$ ) on chromosome 13 (Figure 58) (Results also summarised in Table 11).

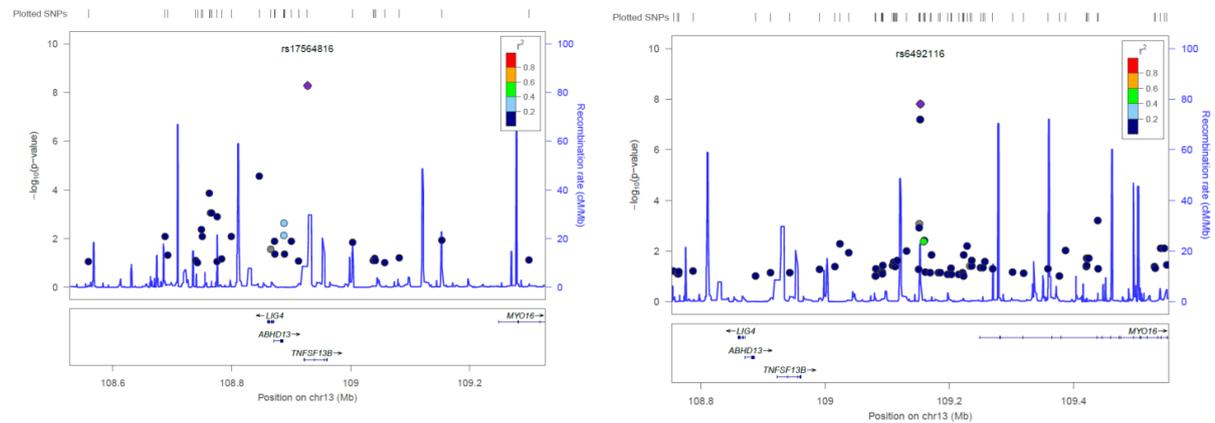


**Figure 56** – Regional association plot for EBNA-1 and rs1317155 (in CLSTN2) in YRI. Genomic location indicated by the X-axis, P-value by the Y-axis SNP with lowest P-value indicated by a purple diamond;  $r^2$  scale showing the degree of LD between the lowest P-value SNP other SNPs displayed in the upper left hand corner; only associations at  $p < 1E^{-01}$  are shown.



**Figure 57** - Regional association plot for EBNA-1 and rs17276586 (CDH13) in YRI. Only associations with  $p < 1E^{-01}$  are shown.

rs17564816 was located within *TNFSF13B* and showed association to EBNA1 levels in YRI. Also in CEPH samples three SNPs located immediately downstream of the gene showed significant association, but to EBNA2 transcript variation (Figure 58).



**Figure 58** – Regional association plot for EBNA-1 and rs17564816 in YRI samples (left) and EBNA-2 and rs6492116 in CEPH samples (right). Only associations with  $p < 1E^{-01}$  are shown.

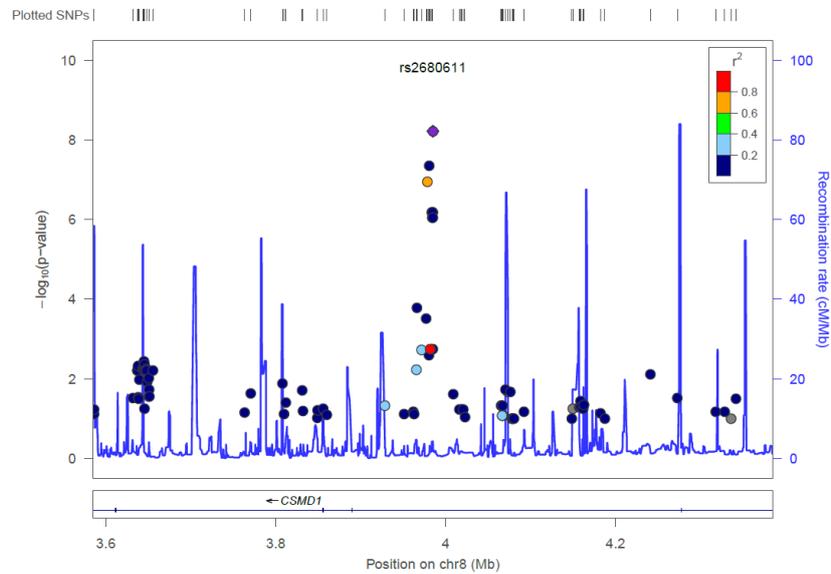
#### YORUBA Panel LCLs

SNP	Trait	P-value	Chr	BP	Gene	Left Gene	Right Gene
rs16849795	EBNA-1	8.84E <sup>-08</sup>	3	1.41E <sup>08</sup>	CLSTN2	NMNAT3	TRIM42
rs1317155	EBNA-1	1.92E <sup>-08</sup>	3	1.41E <sup>08</sup>	CLSTN2	NMNAT3	TRIM42
rs16849869	EBNA-1	6.51E <sup>-08</sup>	3	1.41E <sup>08</sup>	CLSTN2	NMNAT3	TRIM42
rs17564816	EBNA-1	5.23E <sup>-09</sup>	13	1.08E <sup>08</sup>	TNFSF13B	ABHD13	MYO16
rs17276586	EBNA-1	4.22E <sup>-08</sup>	16	81370274	CDH13	MPHOSPH6	HSBP1
rs17192152	EBNA-1	7.40E <sup>-08</sup>	16	81376617	CDH13	MPHOSPH6	HSBP1
rs8051326	EBNA-1	7.40E <sup>-08</sup>	16	81380213	CDH13	MPHOSPH6	HSBP1

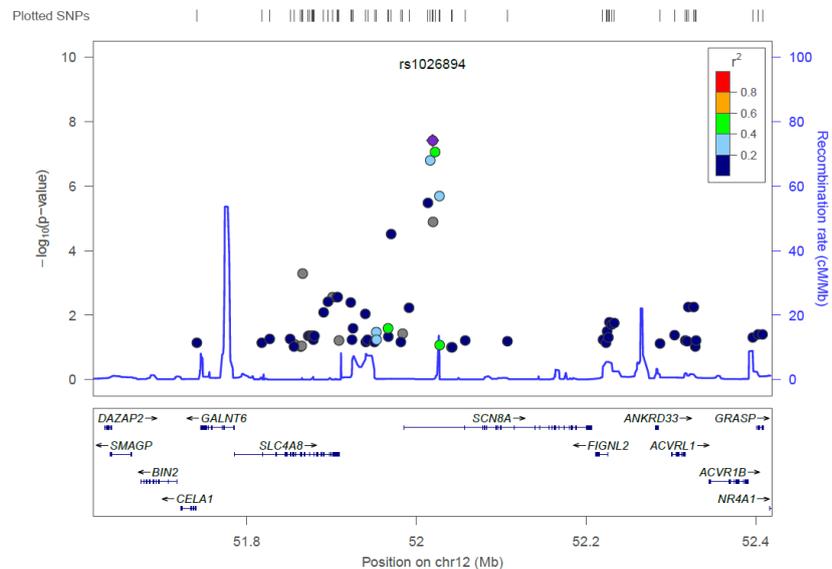
**Table 11** - EBV eQTL results in the RNA-seq samples. Table presents most significant associations found at  $p < 1E^{-07}$ . SNPs located no further away from each other than 200kb are color-coded.

- i) SNP – statistically significant associations are listed by SNP in the first column.
- ii) Trait – lists the EBV latency transcript associated with a relevant SNP
- iii) P-value – reports the P-value of the test statistic for a relevant association test
- iv) Chr – gives the genomic location of a significant association by chromosome
- v) BP – gives the genomic location of a significant association by base pair number
- vi) Gene – reports the genomic location of the associated SNP by gene
- vii) Left/Right Gene – lists the two closest genes flanking a given SNP

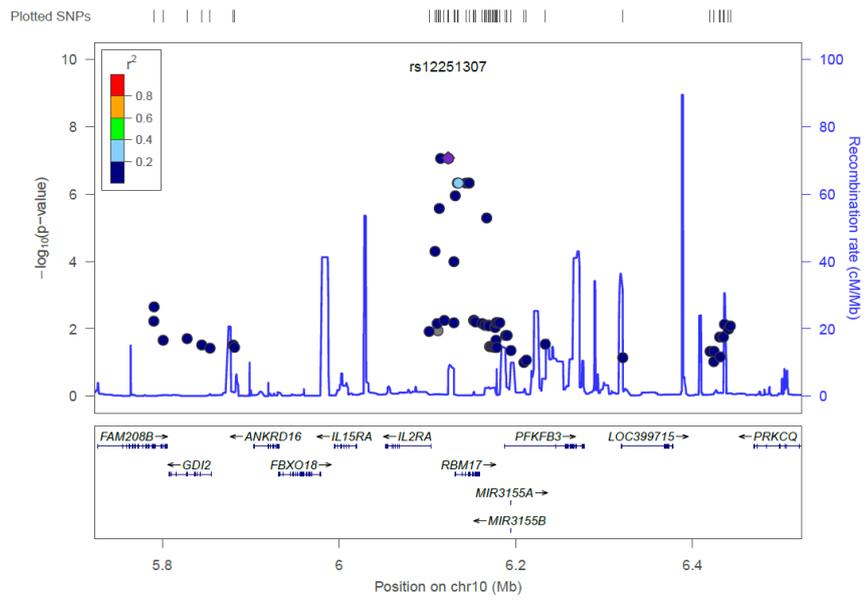
In the CEPH panel, a number of significant associations were found (summarised in Table 12), including for EBER1 (Figure 59-61), EBNA-LP (Figure 62-63) and EBNA2 (Figure 64). Some identified genomic loci showed association with more than one EBV transcript.



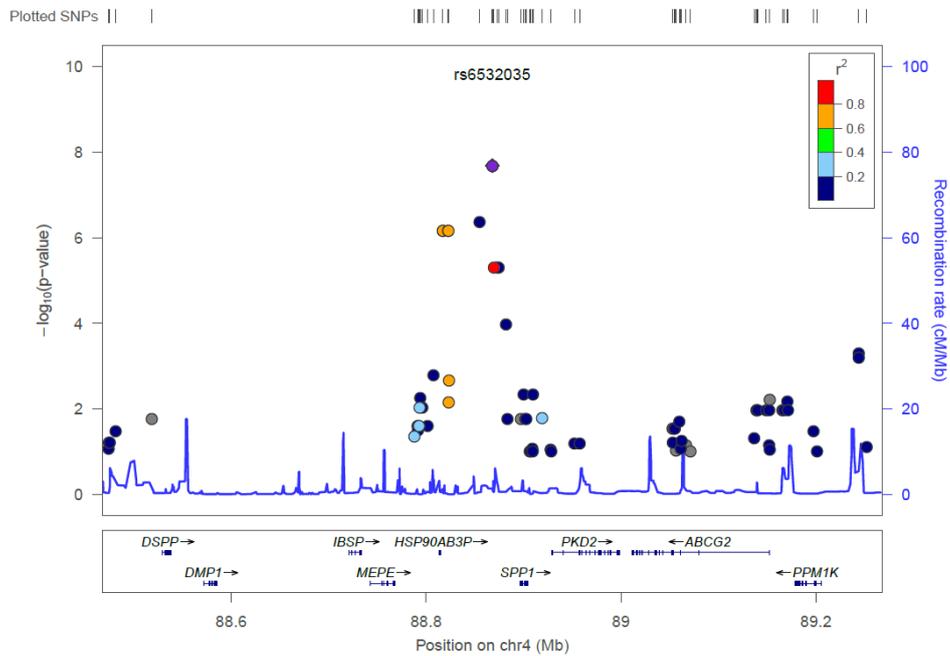
**Figure 59** - LocusZoom regional association plot for CSMD1 and EBER1 in CEPH samples (SNP with lowest P-value indicated by purple diamond;  $r^2$  scale indicating the degree of LD between lowest P-value SNP and others in the upper right hand corner). Only associations with  $p < 1E^{-01}$  are shown.



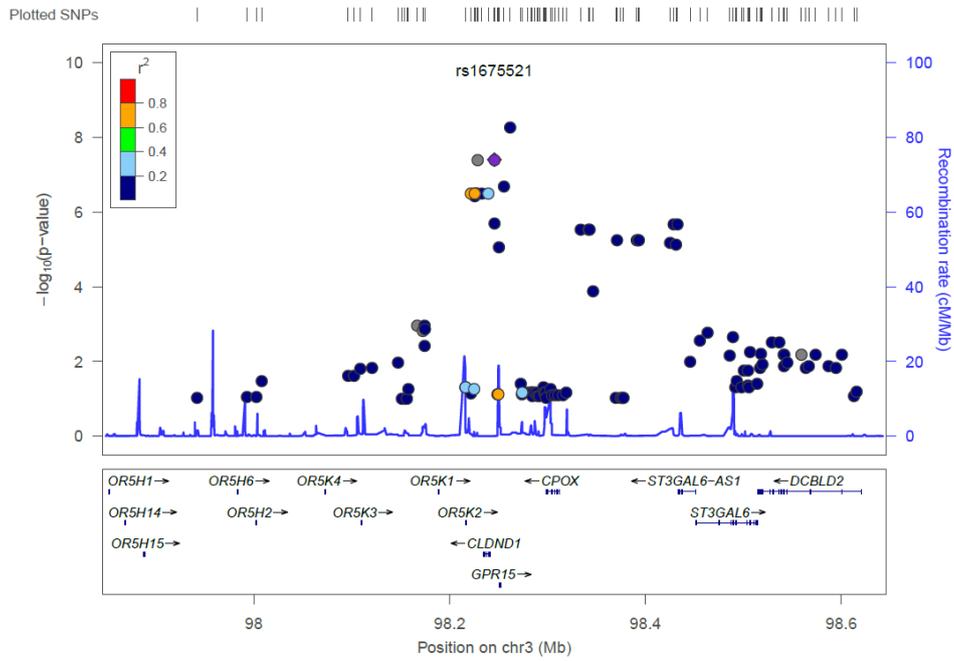
**Figure 60** – Regional association plot for EBER1 at rs 1026894 (within SCN8A). Only associations with  $p < 1E^{-01}$  are shown.



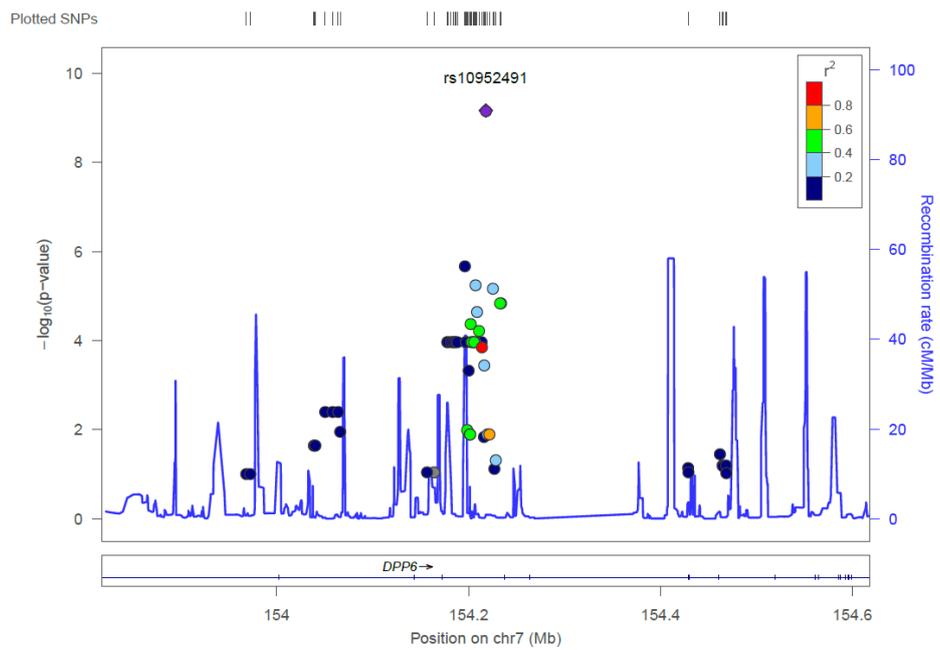
**Figure 61** – Regional associations of EBER1 and rs12251307 (within IL2RA) and proximal SNPs. Only associations with  $p < 1E^{-01}$  are shown.



**Figure 62** – EBNA-LP and rs6532035 (SPP1) association in CEPH. Only associations with  $p < 1E^{-01}$  are shown.



**Figure 63** - Regional association plot for rs1675521 (within GPR15) and EBNA-LP in CEPH samples; only associations with  $p < 1E^{-01}$  are shown.



**Figure 64** - Association between EBNA-2 and rs10952491 locus (within DPP6). Only associations with  $p < 1E^{-01}$  are shown.

CEPH Panel LCLs							
SNP	Trait	P-Value	Chr	BP	Gene	Left Gene	Right Gene
rs6788120	EBNA-LP	3.97E-08	3	99711169		OR5K2	CLDND1
rs1675521	EBNA-LP	3.97E-08	3	99728322		LOC10013	GPR15
rs1350790	EBNA-1	6.02E-08	3	99738082		GPR15	CPOX
rs9784358	EBNA-1	2.76E-08	3	99744348		GPR15	CPOX
rs9784358	EBNA-LP	5.54E-09	3	99744348		GPR15	CPOX
rs6532035	EBNA-LP	2.07E-08	4	89086985		HSP90AB3	SPP1
rs10029763	EBNA-LP	2.07E-08	4	89087020		HSP90AB3	SPP1
rs796398	EBNA-1	8.24E-09	6	83113039		LOC10013	TPBG
rs796398	LMP-2A	9.28E-08	6	83113039		LOC10013	TPBG
rs770898	EBNA-1	2.71E-10	6	83122607		LOC10013	TPBG
rs770898	LMP-1	5.31E-08	6	83122607		LOC10013	TPBG
rs770898	LMP-2A	3.16E-08	6	83122607		LOC10013	TPBG
rs6930908	EBNA-1	2.71E-10	6	83123018		LOC10013	TPBG
rs6930908	LMP-1	5.31E-08	6	83123018		LOC10013	TPBG
rs6930908	LMP-2A	3.16E-08	6	83123018		LOC10013	TPBG
rs1570140	EBNA-1	2.71E-10	6	83129590	TPBG	LOC10013	TPBG
rs1570140	LMP-1	5.31E-08	6	83129590	TPBG	LOC10013	TPBG
rs1570140	LMP-2A	3.16E-08	6	83129590	TPBG	LOC10013	TPBG
rs9361923	EBNA-1	2.71E-10	6	83172329		TPBG	UBE2CBP
rs9361923	LMP-1	5.31E-08	6	83172329		TPBG	UBE2CBP
rs9361923	LMP-2A	3.16E-08	6	83172329		TPBG	UBE2CBP
rs2208422	EBNA-1	6.47E-10	6	83175011		TPBG	UBE2CBP
rs2208422	EBNA-LP	6.85E-08	6	83175011		TPBG	UBE2CBP
rs3011889	EBNA-1	2.87E-10	6	83194997		TPBG	UBE2CBP
rs3011889	LMP-1	5.98E-08	6	83194997		TPBG	UBE2CBP
rs3011889	LMP-2A	3.74E-08	6	83194997		TPBG	UBE2CBP
rs10952491	EBNA-1	6.89E-10	7	1.54E+08	DPP6	FLJ16734	DPP6
rs10952491	EBNA-LP	9.17E-11	7	1.54E+08	DPP6	FLJ16734	DPP6
rs17404676	EBER1	4.41E-08	8	3968021	CSMD1	MYOM2	LOC780813
rs2680611	EBER1	6.06E-09	8	3972317	CSMD1	MYOM2	LOC780813
rs10905718	EBER1	8.38E-08	10	6154862		IL2RA	RBM17
rs12251307	EBER1	8.38E-08	10	6163501		IL2RA	RBM17
rs7100400	EBER1	8.38E-08	10	6164086		IL2RA	RBM17
rs1026894	EBER1	3.84E-08	12	50305426	SCN8A	SLC4A8	LOC728522
rs17125925	EBER1	8.45E-08	12	50308163	SCN8A	SLC4A8	LOC728522
rs1886195	EBNA-2	6.18E-08	13	1.08E+08		TNFSF13B	MYO16
rs1886197	EBNA-2	6.18E-08	13	1.08E+08		TNFSF13B	MYO16
rs6492116	EBNA-2	1.58E-08	13	1.08E+08		TNFSF13B	MYO16

**Table 12** – EBV latency eQTLs in the RNA-seq samples (at  $p < 1E^{-07}$ ).

- i) SNP – statistically significant associations are listed by SNP in the first column.
- ii) Trait – lists the EBV latency transcript associated with a relevant SNP
- iii) P-value – reports the P-value of the test statistic for a relevant association test
- iv) Chr – gives the genomic location of a significant association by chromosome
- v) BP – gives the genomic location of a significant association by base pair number
- vi) Gene – reports the genomic location of the associated SNP by gene
- vii) Left/Right Gene – lists the two closest genes flanking a given SNP

## **3.7 Discussion**

### **3.7.1. LMP1A eQTL**

The results of the experiment indicated no significant genetic regulatory effects of the SNPs within the tested region on the expression of 4 local genes. This was also the case with the candidate LMP1 eQTL (rs1913243). Only for CCR8, in samples derived from both normoxic and hypoxic conditions, did the putative eQTL come up as weakly associated with transcript abundance. The eQTL was in LD with some of the neighbouring SNPs however all had relatively low P-values. This might be indicative of association and worth further investigation using a larger sample and a denser SNP map updated from the 1000 Genomes project. Particularly that the association between changes in expression of some members of the *CCR* and *CXCR* family and EBV infection in B-cell lymphomas has been already observed (Nakayama 2002, Benned-Jensen et al. 2011, Deutsch et al. 2008).

### **3.7.2. Latency eQTLs in the hypoxia study panel**

After quantifying 10 latency transcripts in the 58 YRI LCLs from the hypoxia study and testing for genome-wide association, a single eQTL candidate was identified significant for multiple latency phenotypes. The most consistent and statistically significant peak identified was within the *SSH2* on chromosome 17. The peak consisted of multiple SNPs most of which were in high linkage disequilibrium. Haplotype and LD structure analysis could be conducted to increase the power to detect potential causal allele, however the data may also need a more stringent normalisation than by using delta-Ct values on their own. Figures 24-25,34-25 and 37-38 depict the associations as regional association plots and box plots. Genotypes of

multiple SNPs at this locus correlated with the EBER1, EBNA3C and LMP1 transcript variability.

The *SHH2* association can be explained by comparisons of levels of expression of the three genes grouped according to the genotype for the lowest P-value SNP. The associated sequence on chromosome 17 encompasses a broad fragment with several genes of interest and markers implicated in height and platelet volume by recent GWAS studies (Lango et al. 2010 Soranzo et al. 2009). *SSH2*, located centrally, belongs to the fairly recently discovered family of slingshot phosphatases (Jung et al. 2007). It is involved in dephosphorylation of cofilin, which changes actin cytoskeleton during mitosis (Kligys et al. 2007). Slingshot phosphatases may therefore play a critical role in animal cytokinesis (Kaji et al. 2003). *CCDC55* is a recently described novel mRNA splicing regulator (Kim et al. 2011). Like *SSH2*, *CCDC55* and its neighbour *EFCAB5* have been recently annotated and their function has still not been well described (Cheung et al. 2012 Roehl et al. 2010). Their sequences appear to be conserved across mammalian species and form an extended run of homozygosity (Cheung et al. 2012 Roehl et al. 2010). *GIT1* is an intracellular scaffolding protein that interacts with a range of proteins including those involved in the MAP kinase pathway, such as MEK1 and ERK1/2 and is involved in diverse processes, including agonist-coupled receptor endocytosis and focal adhesion assembly (Hartford et al. 2009). Consequently the genes of interest are either poorly described or have broad functions and there seems to be no specific functional background which could link these proteins to particular viral latency transcripts. Association between EBNA-2 and SNPs located proximal to *ASAPI* appears interesting because of the MS context, however lacks a discrete association peak.

Among other functions, *ASAPI* has been implicated in remodelling of actin cytoskeleton (Randazzo et al. 2000, Bharti et al. 2007). This functional background is also shared by genes

associated with other candidate eQTLs, namely *SSH2* and *GIT1* (Kligys et al. 2007, Turner et al. 2001).

The Mohr study LMP1 eQTL SNP rs1913243 was not significantly associated with expression levels of the latency transcripts tested.

### **3.7.3 Latency eQTLs in the MRC-A panel**

Out of the significant candidates with MAF exceeding 0.05, eQTL in the region of the *RASGEF1A* and *FKBP6* genes appeared most interesting, especially as the SCAN database identified the associated SNPs as eQTLs for genes with risk loci to EBV-related autoimmune diseases (SLE, MS) or lymphoma-relevant genes. This appeared unusual given the small number of significant associations.

*RASGEF1A*, expressed mostly in CNS, is involved in G protein signalling and transcriptional regulation and has been reported to interact with K-RAS, H-RAS, and N-RAS signal transduction in vitro, which might indicate it upregulates the Ras GTPases (also modulated by LMP2A) and thus influences cellular growth and survival. In vitro suppression of *RASGEF1A* by small interfering RNA (siRNA) slowed the development of cholangiocarcinoma (Ura et al. 2006). Gain of *RASGEF1A* and its overexpression was also reported in cases of testicular embryonal carcinoma (Gilbert et al. 2011). More importantly one of the SNPs most significantly associated with EBNA1 levels, has previously been identified as an eQTL for CD83 cell surface receptor and a member of the immunoglobulin superfamily known to be involved in EBV transformation and upregulated by viral LMP1 oncogene in infected B-cells (lymphoblasts) (Dudziak et al. 2003). CD83 is a marker of mature dendritic cells but it is also expressed by activated lymphocytes, mononuclear cells

and neutrophils (Dudziak et al. 2003, Pan et al. 2006, Ehlers et al. 2013). It is thought to have pleiotropic effects, including activation of naïve T-cells and helper T- cells, naïve B-cell activation, and play important role as one of immune response regulators, however its precise functions, ligands and transcriptional regulation remain unknown (Ehlers et al. 2013, Stein et al. 2013).

Another significant association, rs17339199, was located within *FKBP6*, and correlated with EBNA3A transcript levels. FKBP belongs to the family of immunophilins involved in protein folding and cell signaling, and homologous chromosome pairing in meiosis during spermatogenesis (Crackower et al. 2003). It is expressed in all tissues, but present at highest level in testis, liver, kidney, skeletal muscle and heart. Mutations in this gene have been associated with male infertility in humans (Zhang et al. 2007). FKBP6 is one of 25-28 genes deleted in Williams syndrome. Its methylation has been reported to be higher in malignant cases of cervical cancer and also to correlate with the stage of breast cancer (Fujikane et al. 2010, Brebi et al. 2013). SCAN indicated that the SNP is also eQTL for proinflammatory IL-12B. IL-12B is produced by immune system's non-T antigen presenting cells, encodes a p40 subunit common to IL-12 and IL-23 and stimulates the differentiation of T helper cells and secretion of interferon- $\gamma$  (IFN- $\gamma$ ) by T and natural killer (NK) cells (Yilmaz and Yentur 2005). This gene has been associated with a wide range of diseases. Polymorphisms in IL12-B causing lower expression have been reported to correlate with disease outcome in hepatitis C virus infection (Houldsworth et al. 2005). Variants within the gene were also associated with risk of breast cancer NPC and MS (Wei et al. 2009, Kaarvatn et al. 2012, Varade et al. 2012, Sawcer et al. 2011). Also variants within IL-12A were confirmed to be associated with higher risk of MS (Sawcer et al. 2011). Cytokine gene polymorphisms, for instance IL-10, are associated with susceptibility to, and reactivation of EBV (Nakai et al. 2002). Different levels of IL-12 expression were linked to increased lytic replication in LCLs (Arvey et al. 2012).

Higher IL-12 levels were reported in the sera of tonsillar tissues of IM patients (Nakai et al. 2002). Also EBV DNA load in serially collected PBMCs from patients with primary EBV infection correlated with levels of IL12 in several cases, which indicated that it might be influencing the kinetics of EBV DNA load in IM patients (Nakai et al. 2002). Significant differences in IL-12B expression were observed in Hodgkin lymphoma patients and their healthy co-twins (Cozen et al. 2009). Overall, the joint EBNA3A and IL12-B eQTL appears to be an interesting candidate particularly because of direct links to EBV infection, B-cell immune response as well as presence of risk loci for autoimmune and lymphoid disorders. Unfortunately, the association consisted of a single solitary SNP.

The third candidate, rs12660137, was located approximately 20-25kb in between *C6orf52* and *GCNT2*, was found to be associated with EBER-1 transcript levels. A search through public eQTL databases indicated that the SNP has also been associated with the tumour necrosis factor receptor superfamily member 1B (or TNFRSF1B) expression levels. This tumour necrosis factor receptor mediates the effects of TNF $\alpha$  signaling by activating intracellular NF- $\kappa$ B (similarly to LMP1 signalling) and multiple inflammatory pathways (Albarga 2002, Medrano et al. 2014). Polymorphisms in this receptor may be affecting the binding efficiency of TNF and affect its outcome (Medrano et al. 2014). Most interestingly a polymorphism in this gene has been associated with SLE risk in a cohort of 81 Japanese patients with SLE and 207 healthy individuals by candidate-gene study (Tsuchiya et al. 2001). This association has not been replicated by a more case-control candidate-gene study relying on Caucasian cohorts (Tsuchiya et al. 2001). Another candidate-gene case control study conducted in 1006 White patients with Coronary Artery disease and 183 healthy controls found significant association for a TNFRSF1B marker ( P-value of 0.00000069) concluding that genetic variation within the gene may predispose to the disease (Benjafeld et al. 2001). Finally, a similar study in 42 patients with Polycystic Ovary syndrome and 36

controls from Spain claimed an association with a CA-repeat microsatellite polymorphism within TNFRSF1B (P-value < 0.03; Peral et al. 2002). Additionally a study which applied pathway analysis to SNPs derived from a publicly available MS GWAS dataset ([NCBI dbGap at http://www.ncbi.nlm.nih.gov/gap](http://www.ncbi.nlm.nih.gov/gap)) by integrating linkage disequilibrium (LD) analysis, functional SNP annotation and pathway-based analysis (PBA), has proposed a possible link to Multiple Sclerosis (Song et al. 2013). Also polymorphisms in TNFRSF1A have been associated with MS risk (Sawcer et al. 2011), and variation in another protein from the same family, TNFRSF19, was linked to NPC risk (Hildesheim and Wang 2012).

Another suggestive association for the EBER1 transcript level was present on chromosome 16 and resulted in a clear association peak within the ORF of the RNA binding protein 1 *RBFOX1* – a protein moderating neuronal excitation in mammalian brain (Gehman et al. 2011). The RNA binding proteins regulate alternative splicing in the nervous system however little is known about RBFOX1 interactions (Weyn-Vanhentenryck et al. 2014). Disruption of *Rbfox1* has been implicated in autism (Weyn-Vanhentenryck et al. 2014). A suggestive association for EBER2 was in *DCC*, a tumour suppressor gene deleted in colorectal and gastric cancers (Kataoka et al. 2000). Interestingly the most significant SNP within the peak has been indicated by the eQTL databases to be associated with transcription of over 10 human genes including *SMC5*, participating in DNA break repair (Gallego-Paez et al. 2014).

#### **3.7.4 Latency eQTLs in the RNA-seq panels**

Analysis of the latency transcripts quantified by Arvey et al. (2012) in Caucasians from Utah (CEPH) and Yoruba Africans yielded a number of significant associations. This number was even higher than in case of MRCA panel, most likely because no normalisation had been applied. For EBNA1, associations in YRI involved SNPs in the region of *CLSTN2*, *CDH13*

and *TNFSF13B*. Some regions harboured significant associations in both the Yoruban and CEPH. These included SNPs located within or proximally to *TNFSF13B*, as well as *CSMD1* although in the latter case a solitary lytic transcript association was present in YRI samples.

*CLSTN2* encodes a synaptic protein called calsynenin 2, which is predominantly expressed in the CNS as components of the postsynaptic membrane and play a role in postsynaptic signalling (Preuschhof et al. 2010); while *CDH13* encodes cadherin-13, a member of the cadherin family proteins, which span the membrane enabling cell adhesion. Non-synonymous mutations of *CLSTN2* has been reported for cases of gastric cancer (Majewski et al. 2013); however there has been no causal involvement proposed. The gene is predominantly expressed in the brain however also in bone marrow, tonsils and in a Burkitt's lymphoma Daudi cell line, and thus it has been suggested it might have an immunological function (Mero et al. 2013).

Methylation or decreased expression of *CDH13* was shown in cases of solid tissue cancer including breast, lung, colorectal, cutaneous squamous cell carcinoma and NPC (Sun et al. 2007 Takeuchi et al. 2002 Toyooka et al. 2002), and it has been suggested it may act as tumour suppressor (Toyooka et al. 2002). The gene is silenced by LMP1 signalling in NPC cell lines through methyltransferases *DNMT1*, *DNMT3A*, and *DNMT3B* and a similar phenomenon has been observed in endothelial cells infected with KSHV (Leonard et al. 2014, Leonard et al. 2012, Paschos and Allday 2010). However there is no direct link between *CDH13* and disorders of the lymphatic system or EBV infection.

Solitary associations were present within *TNFSF13B*, a member of the tumor necrosis factor (TNF) superfamily. This gene encodes a ligand, promoting B-proliferation and facilitates the activation of immune response. It is involved in BCR signalling and therefore plays a role in the vital pathway hijacked by LMP2 oncogene (Vockerodt et al. 2013). *TNFSF13B* acts as a

ligand for three TNF-like receptors which are present on B cells (Salzer et al. 2005). Its polymorphisms have been associated with both autoimmune (RA and SLE, common variable immunodeficiency) and lymphatic disorders (B-cell non-Hodgkin's lymphoma) (Kawasaki et al. 2002, Salzer et al. 2005). Transgenic mice models harbouring *TNFSF13B* mutations have elevated levels of mature B and effector T cells, and develop symptoms resembling autoimmune disease (Mackay et al. 2009). Also, high levels of the ligand are associated with chronic lymphocytic B-cell leukemia and polymorphisms in *TNFSF13B* promoter (Novak et al. 2009). This makes it an interesting candidate since it is connected to the broad scope of disorders in which EBV plays a role of a co-factor. In addition, its function is utilised by the virus during transformation (Latency III) and it is upregulated by the EBNA3C latency protein (Zhao 2011). The associated SNPs clustered either in the middle intronic part of the gene's ORF (in Yoruban samples) and reveal correlation to EBNA-1 levels, or (in CEPH samples) are located less than 200kb downstream of the ORF forming a peak of 3 SNPs above the suggestive threshold of  $1 \times 10^{-5}$ . In CEPH they are correlated with the levels of EBNA-2 (but also a lytic protein BHRF1).

Association tests in the Caucasian LCLs resulted in 7 other significant associations. One of them involved SNPs correlated with EBER1 levels, located within *CSMD1*, CUB and Sushi multiple domains-1, expressed in brain and epithelial tissues (Kraus et al. 2006). This locus also appeared to be significantly associated to a lytic transcript, BZLF1, in both the CEPH and the Yoruban individuals (in addition to EBER1 in the latter). Only a solitary SNP, downstream of *CSMD1*, appeared significantly associated with BZLF1 levels in the Yoruban LCLs, while three SNPs associated with BZLF1 in CEPH samples were located within the ORF of *CSMD1*. *CSMD1*, has been linked to a range of disorders including schizophrenia (Rose et al. 2013, Steen et al. 2013, Donhoe 2013) and MS (Baranzini et al. 2009). Its link to MS however has not been replicated in the more recent study (Sawcer et al. 2011). This is a

recently described gene whose function is poorly understood, but it is thought to play a role in neurodevelopmental disorders affecting cognition (Rose et al. 2013) but has also been linked to cancer (Shull et al. 2013). Only two studies investigated *CSMD1* function, suggesting it may act as tumour suppressor or play a role in the immune system's complement cascade by tagging pathogens for elimination (Distler et al. 2012, Havik et al. 2011). It is worth noting that *CSMD1* and proximal sequences contain over significant 20 markers from multiple GWA studies.

In CEPH LCLs, an interesting significant association was found between *EBER1* and rs12251307 (P-value 8.38E-08) (Figure 61). The same variant has been found to be associated with the risk of type 1 diabetes in a case-control study based on 7,514 cases and 9,045 controls (P-value 1E-13) and another one with 3,561 cases, 4,646 controls (Barrett et al. 2009 Cooper et al. 2008). The reported gene, *IL2RA*, encodes interleukin 2 receptor alpha displayed on the surface of immune cells. Interestingly variants of the gene have been associated with increased risk of MS by the most recent GWAS (Sawcer et al. 2011) and replicated in an independent study (Rubio et al. 2008). Very recently, another study reported that variants in *IL2RA* increase the risk of MS even further when combined with Human herpesvirus 6 (HHV-6) infection and specific HLA haplotypes (Rahal et al. 2014). One of the MS-associated SNPs, rs7090512, is located approximately 10kb away and also downstream of the gene. However it is more proximal to its ORF. The other is further upstream, within *IL2RA*'s ORF, close to other markers associated with increased risk of RA (Stahl et al. 2010, Orozco et al. 2014). The numerous associations highlight important role of the gene in immune response. Interleukin 2 stimulates T-cell proliferation, survival and differentiation and its receptor is an important marker of CD4+ CD25+ regulatory T cells, characterised by FoxP3 expression, that ward the organism against autoimmunity (Takahashi et al. 2002).

In CEPH samples the association test revealed two candidate EBNA-LP eQTLs located close to *SPPI* (Figure 62), a protein known to upregulate the viral Cp/Wp promoters and speculated to bind EBNA1 (Arvey et al 2012; Lu et al 2011). A potential EBNA1 eQTL was located close to *GPR* (Figure 63), one of co-receptors which can be utilised by human and simian immunodeficiency viruses for entry, along with other transmembrane receptors, including *CCR* and *CXCR* family (Titti et al. 2002). Also SNPs in *SCN8A* (Figure 60) correlated with EBER1 levels. *SCN8A* is a gene expressed in the CNS and encodes a sodium-channel structural protein whose mutant has been implicated in epileptic encephalopathy (Burgess et al. 1995, Veeramah et al. 2012). Among other associations found in CEPH were variants in *DPP6* (Figure 64). Both EBNA-LP and EBNA2 transcripts were significantly correlated with the genotype of a single SNP in *DPP6*. *DPP6* is expressed in brain and its product, Dipeptidyl aminopeptidase-like protein 6 is responsible for modulating the biophysical properties of voltage-gated potassium channels (van Es et al. 2008). Variants in the gene have been associated to susceptibility to amyotrophic lateral sclerosis, a neurodegenerative disorder, in European populations (van Es et al. 2008).

In CEPH samples, four transcripts EBNA1 EBNA-LP LMP1 and LMP2A were correlated with SNPs within and proximal to *TPBG* on chromosome 6 (Appendix Figure A4). *TPBG*, or trophoblast glycoprotein (also known as 5T4) is a transmembrane protein whose functions are poorly understood, although it is expressed at high levels in placenta and in most common tumours, including 80% kidney, breast, colon, prostate, and ovary carcinomas which makes it a good diagnostic marker (Zhao and Wang 2007). It has also been shown to be significantly upregulated in LCLs (Baik et al. 2007).

### 3.8 Conclusion

In summary, the eQTL assays yielded no consistent genome-wide significant associations. This suggests that the results should be interpreted with caution as the current study's small sample size is a serious limiting factor. Only the familial MRC-A panel samples exceeded 100 individuals. Although family-based cohorts have been suggested as a suitable method of replication for the population-based GWA studies, they provide comparatively less power to detect significant associations (Benyamin et al. 2009).

However several MRC-A-derived associations appear suggestive, in particular SNPs in the region of *RASGEF1A* and *RBFOX1*, linked to EBNA1 and EBER1 respectively, however no immediate relevance to EBV infection and disease is evident. The EBV expression data sourced from the RNA-seq studies provided interesting results. It is worth noting that the two associations to the same loci (although not involving same SNPs) in both CEPH and YRI samples cantered on a gene implicated in immune response (*TNFSF13B*) and another gene previously considered to contain risk locus for an EBV-related autoimmune disease (*CSMD1*). The potentially most interesting finding is the significant association to a type 2 diabetes risk marker downstream of *IL2RA*. The significance of this finding is highlighted particularly by the context of several MS and RA GWAS risk located downstream and within *IL2RA* as well as its function in regulating T-cell immune response and maturation. However, the significance of the findings obtained with the RNA-seq samples is undermined by the lack of normalisation which will be necessary in future meta-analysis.

## **Chapter 4. EBV copy number QTLs**

### **4.1 Introduction**

EBV copy number is an important factor correlated with severity and predicting the outcome in many EBV-positive lymphomas (Hohhaus et al. 2011, Baldanti et al. 2000, Smets et al. 2002, Balandraud et al. 2003). Additionally viral copy number may be directly responsible for the severity of acute and chronic IM (Hadinoto et al. 2008, Maeda et al. 2006, Thorley-Lawson 2001) and therefore may be influencing the course of other lymphoproliferative disorders. Although direct causality for any disorder has not been established, there is evidence that viral load in LCLs is partly genetically determined (Caliskan et al. 2011). A study of viral copy number QTLs and their overlap with known disease risk loci could shed more light on the role of the virus in EBV-positive disease aetiology.

Genetic determinants of EBV viral load were investigated using a publicly available dataset of quantified relative EBV copy number across a panel of 172 LCLs from the HapMap CEPH and Yoruban collections (Choy et al. 2008). In the second part of the experiment, EBV load was quantified across 414 LCLs from the four European Human Variation panels of the 1000 Genomes Project and calibrated using serial dilutions of Namalwa cell line gDNA.

### **4.2 Aims of the chapter**

1. To map EBV copy number as a quantitative trait using samples from the Choy et al. study.
2. To quantify and map EBV copy number using gDNA sourced from the British, Finnish, Iberian and Italian 1000 Genomes LCLs.

### **4.3 Mapping EBV copy number QTLs in HapMap LCLs**

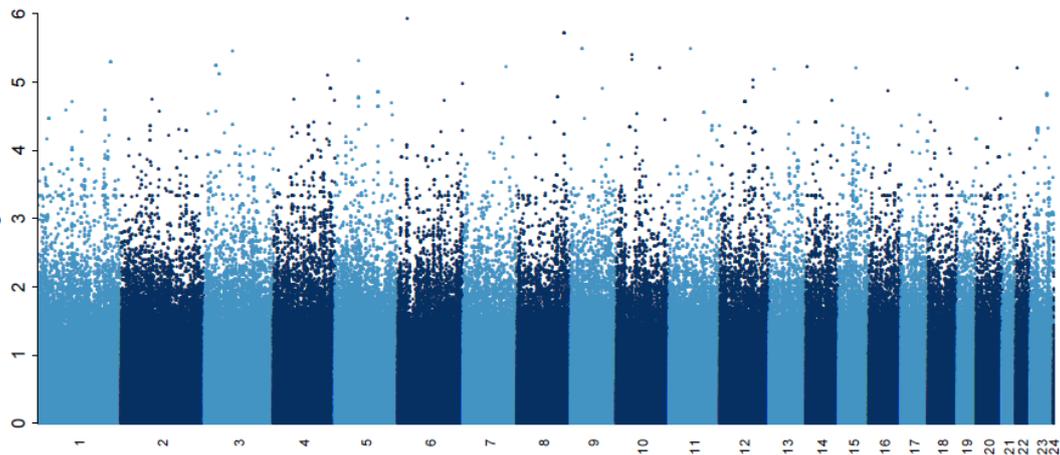
In late 2008 Choy et al. published a study on drug response in a set of Yoruban and Caucasian LCLs from the International HapMap Project (Phase 3 genotypes), in which EBV copy number was quantified as a potential confounder of global gene expression results (Choy et al. 2008). EBV load was determined using custom TaqMan assays targeting a sequence from the viral DNA polymerase gene, and the relative copy number data was made publically available (Choy et al. 2008). The scope of the study did not encompass EBV QTL identification and, consequently, the public EBV expression database provided a means to expand the current search for viral copy number QTLs onto independent cohorts.

#### **Quality Control**

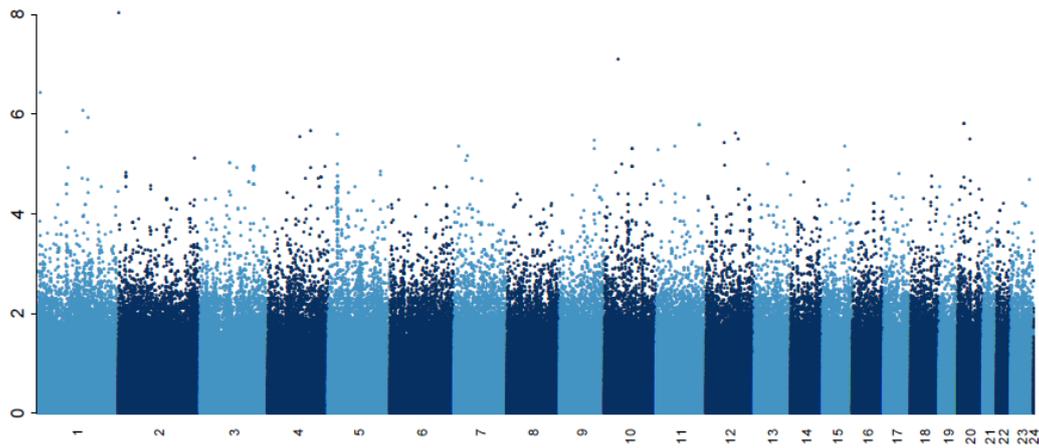
EBV copy number data quantified by Choy et al. (2008) was obtained from Broad Institute's website ([http://www.broad.mit.edu/mpg/pubs/hapmap\\_cell\\_lines/](http://www.broad.mit.edu/mpg/pubs/hapmap_cell_lines/)). The database includes two datasets for the same set of LCLs, each containing log-transformed EBV relative copy values measured at two different stages of the experiment (cell line expansion phase, and main experiment conducted on mature cell lines) as well as a third, smaller dataset sourced from a "replicate" batch of chosen mature LCLs. The EBV copy number has been transformed by Choy et al. (2008) by subtracting the Ct of an internal reference housekeeping gene (NRF1 locus targeted by a 90 bp TaqMan assay) to obtain delta-Ct values (equivalent to log-transformation). Both mature LCL datasets were used in the analysis. According to Choy et al. (2008) the dataset was restricted to (nominally) unrelated individuals only. No further phenotype data transformation was conducted.  $5 \times 10^{-8}$ , a standard genome-wide significance threshold adopted in GWAS studies, was set as the P-value threshold for associations in the current experiment.

## Results

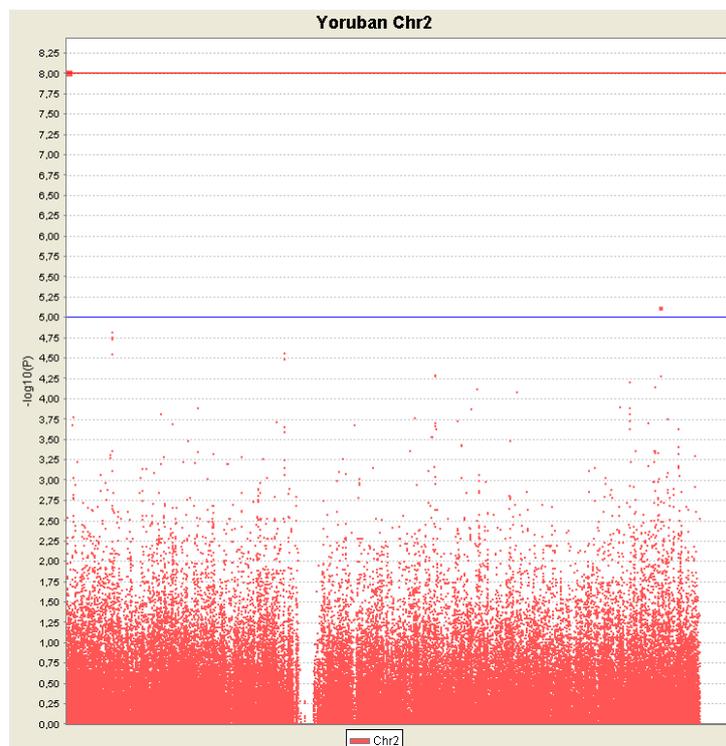
Relative EBV copy numbers were sourced for 73 CEPH Caucasian and 78 Yoruban individuals from the Broad Institute's website and GWAS tests performed using PLINK (--assoc function, MAF>5%, 10% for individual/SNP missing rate, HWE P-value<1x10<sup>-10</sup>). The results are displayed in the plots below (Figures 65-68). Only a single SNP (rs10170606 on Chromosome 2) reached the genome-wide significance threshold with a P-value of 9.35E<sup>-09</sup> in the Yoruban population (Figure 66). The SNP, however, was not part of a larger association peak and is located within an intergenic region on chromosome 2, not immediately associated with any gene (Figure 67). There is no known function of rs10170606. A clear, however statistically not significant, peak was located on chromosome 5 and encompassed *CDH12*, a type II cadherin (Figure 68).



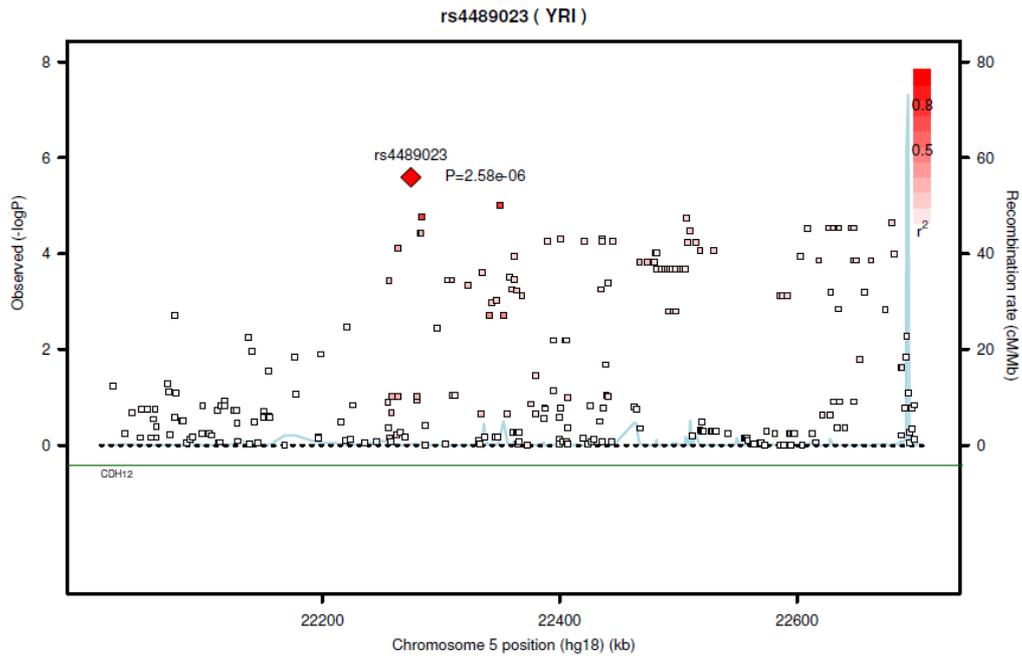
**Figure 65** – EBV copy number whole genome association (73 CEPH Caucasians); Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10}$  P-values displayed on the Y axis.



**Figure 66** - EBV copy number whole genome association (78 Yorubans) Chromosome displayed along the X axis; Chr 23 and 24 denote X and Y chromosomes respectively;  $-\log_{10}$  P-values displayed on the Y axis.



**Figure 67** – EBV copy number whole genome association – HaploView regional plot of chromosome 2 (78 Yoruban cell lines).



**Figure 68** – Regional plot of EBV copy number with rs4489023 (red diamond) and proximal SNPs' (squares) within CDH2 ORF genotypes found in 78 YRI samples. LD indicated by shades of red and r-squared scale in the upper right corner.

The EBV relative copy results were not replicated in the second, smaller replicate dataset from cell lines grown by Choy et al. (2008) for the purpose of intra-experimental copy number variation assessment. No genome-wide significant associations or suggestive association peaks were found (data not shown). Pooling the samples together, although technically not correct, did not provide significant associations either.

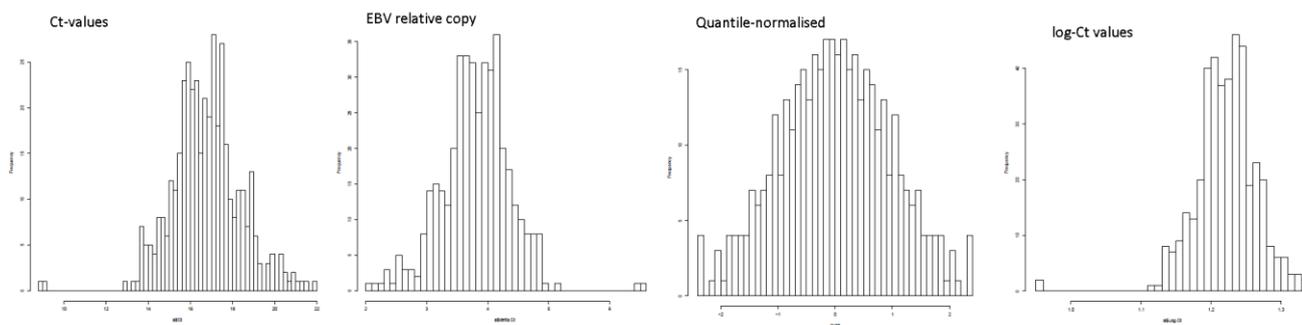
#### 4.4 EBV copy number QTL analysis in 1000 Genomes LCLs

Genomic DNA sourced from 414 LCLs from the 1000 Genomes Project was obtained from Coriell Institute's Cell Repositories and assayed for EBV copy number using an established technique (Caliskan et al. 2011). Briefly, pre-designed TaqMan probes targeting viral IR1 region were used to quantify viral DNA in LCL gDNA samples with concentration normalised to 100ng/ul using Qubit. The relative copy number was then calibrated using a

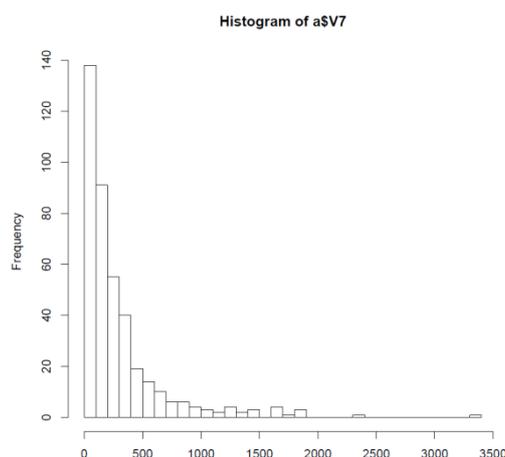
standard curve prepared from serial dilutions of gDNA sourced from the BL Namalwa cell line containing 2 integrated EBV copies per cell (two EBV genomes per single human diploid genome). Namalwa gDNA stock solution was also normalised to a concentration of 100ng/ul using Nanodrop. The tested cohort encompassed all of the European samples from the Human Variation panel representing the Finnish, British, Italian and Iberian populations.

### **Quality Control**

Related individuals including were removed from the analysis manually. The 1000G samples used for the purpose of this study had been provided with the information on cryptic and unexpected relationships, and 2<sup>nd</sup> as well as 3<sup>rd</sup> order relations have been singled out in the 1000G sample spreadsheet available for download from the 1000G website. There were two such pairs (one of a 2<sup>nd</sup> order and one of 3<sup>rd</sup> order) in the current experimental dataset and one individual from each pair was removed from further analysis. Some 1000G genotypes were not available for a subset of Iberian and British samples at the time of the experiment and these samples had to be excluded from the test as well. The qPCR assay had been conducted in duplicate reactions, and whenever the resulting Ct value SD exceeded 1.5 a repeat triplicate reaction was carried out. Samples whose Ct SD exceeded 1.5 repeatedly were removed from further analysis. After QC, 287 samples remained available for statistical tests (over 100 had to be excluded due to missing genotypes at the time). Before the analysis, the relative EBV copy number values were log-transformed and tested for association using PLINK –assoc linear 2 degrees of freedom co-dominant model (MAF > 0.01; 0.1 missing rate, P-val HWE<0.001). Other phenotype data transformations were also considered (Figure 69).



**Figure 69** – Distribution of the EBV copy number shown as (left to right) Ct values, log-relative copy number (gDNA per cell read from the Namalwa gDNA calibration curve), Quantile-normalised relative copy number, delta-Ct values.



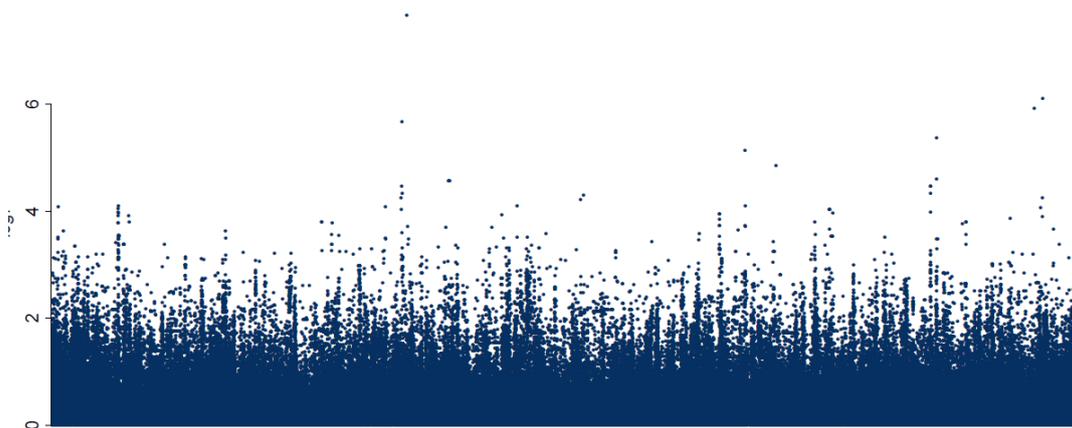
**Figure 69A** – Distribution of the EBV absolute copy number per cell obtained using the method of Caliskan et al. (2011) with Namalwa gDNA calibration curve and conversion factor of 6.6 pg DNA per diploid genome.

The log-transformation was chosen because it provided three interesting candidates, while quantile normalisation generally reduced the overall significance level of all associations to below  $1 \times 10^{-6}$  and  $1 \times 10^{-5}$  genome wide, forcing the association peaks onto the same level as background solitary “noise” associations (Figure A6). Before the tests, PCA was conducted and the test repeated 10 times, each time with a set of 1 to 10 PCs used as covariates as means to control for population structure and potential remaining cryptic relatedness as well as batch effects. No PC correlated with any known variable (population, cDNA plate, qPCR

plate, sex) and using no covariates gave the most significant results, which are summarised in Table 13. As an additional control and to ensure the good quality and consistent levels of the tested gDNA, TaqMan RNase P Detection assay was run on each plate.

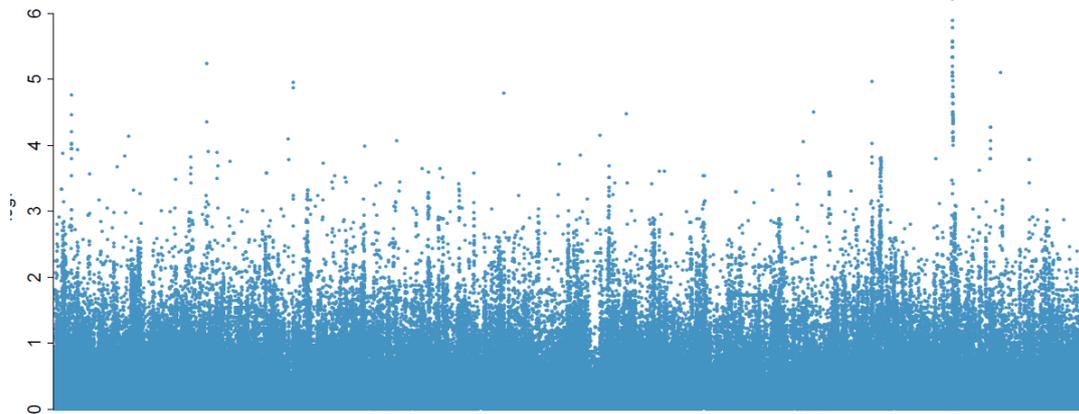
## **Results**

Only a single SNP (rs190636588) on chromosome 14 was associated with viral copy number at genome-wide significance (Figure 70).

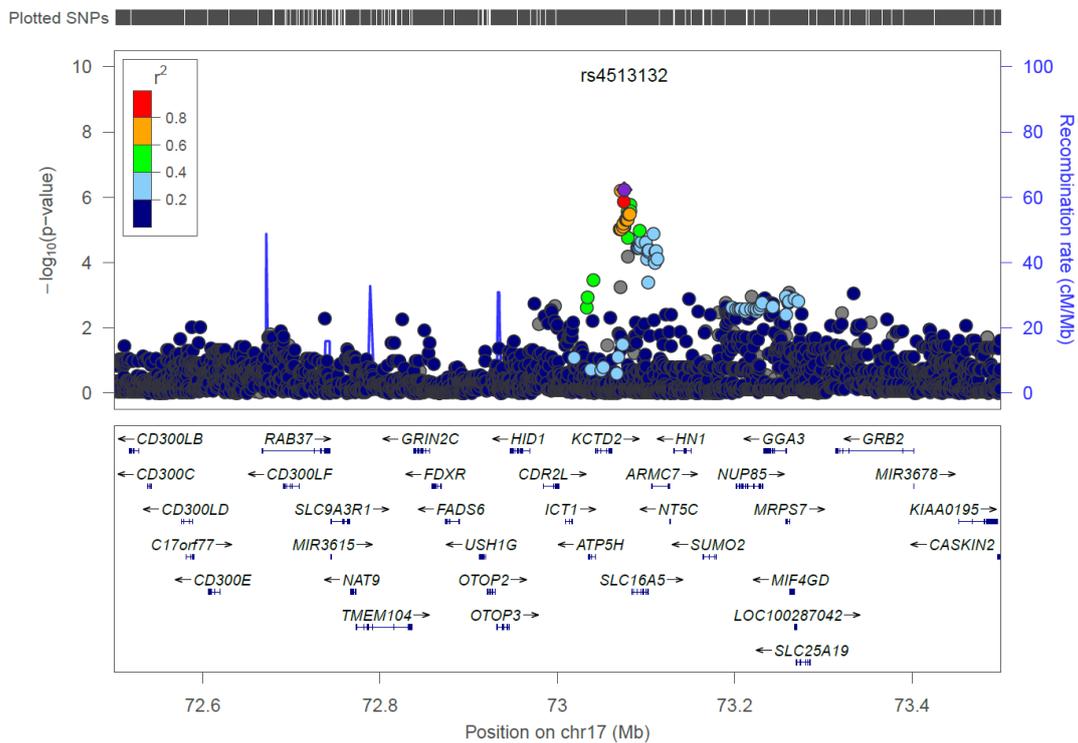


**Figure 70** – Chromosome 14 EBV copy number associations; negative log of the P-value indicated on the X-axis.

rs190636588 is located in a gene desert with the nearest gene *RPS29* located 150kb downstream. A number of other suggestive associations was noted, including two on chromosome 17 and 19 (Figures 71-74). The association peak on chromosome 17 was located within an intergenic region, approximately halfway between the potassium channel tetramerization domain containing 2 (*KCTD2*) ORF and solute carrier family 16 (monocarboxylate transporter) member 5 (*SLC16A5*) gene.

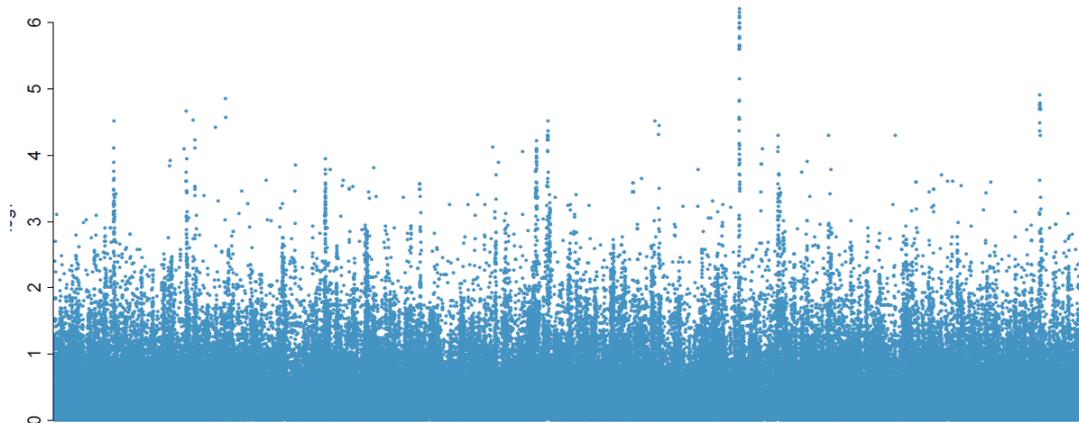


**Figure 71** – Chromosome 17 EBV copy number associations; negative log of the P-value indicated on the X-axis.

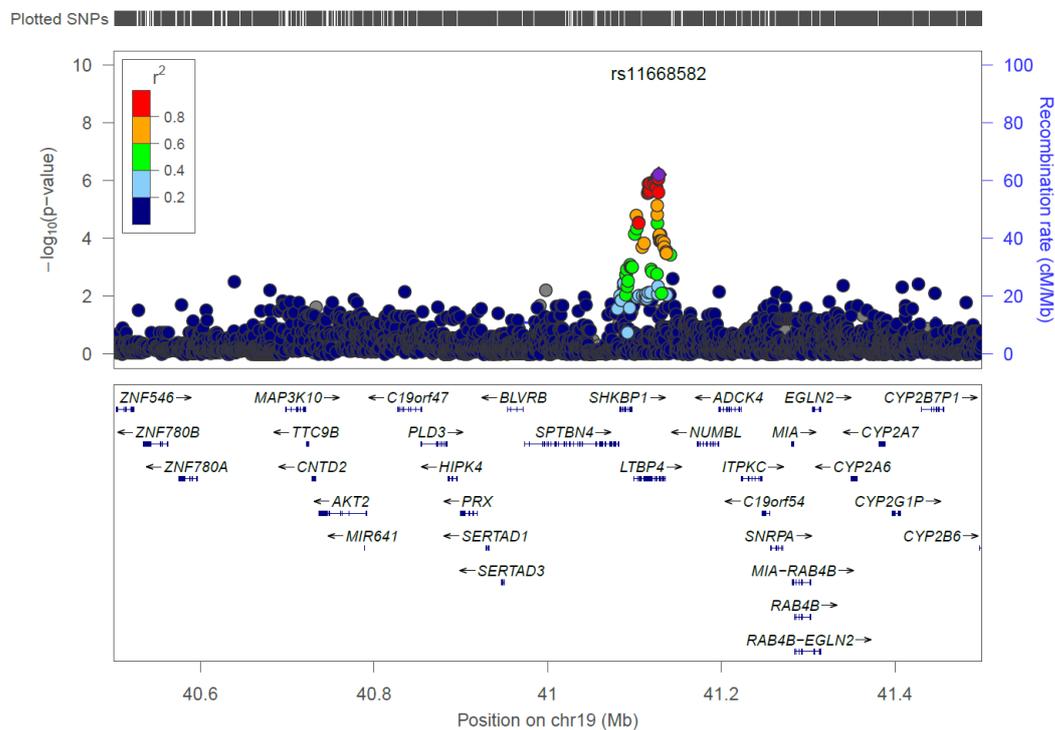


**Figure 72** – Regional association plot showing EBV copy number association with rs4513132 and the SLC16A5 locus.

The association on chromosome 19 involved multiple SNPs located within *LTBP4* gene. *LTBP4* encodes transforming growth factor beta (TGF- $\beta$ ) binding protein.

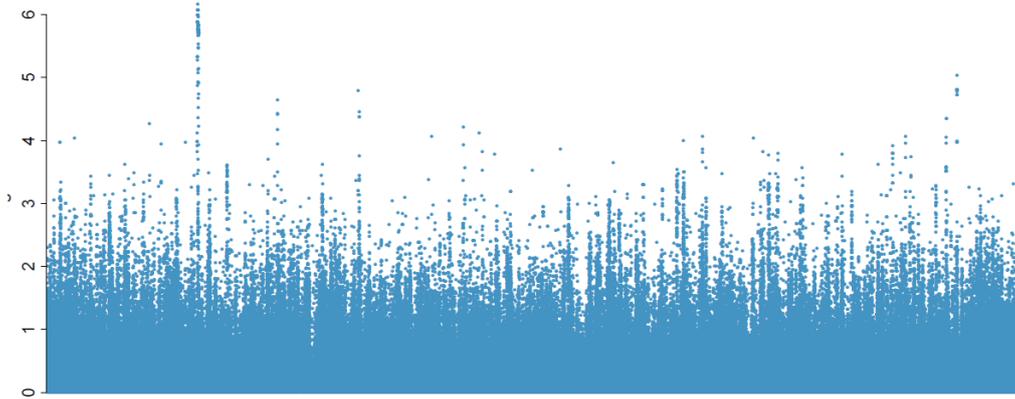


**Figure 73**– Chromosome 19 EBV copy number associations; negative log of the P-value indicated on the X-axis.



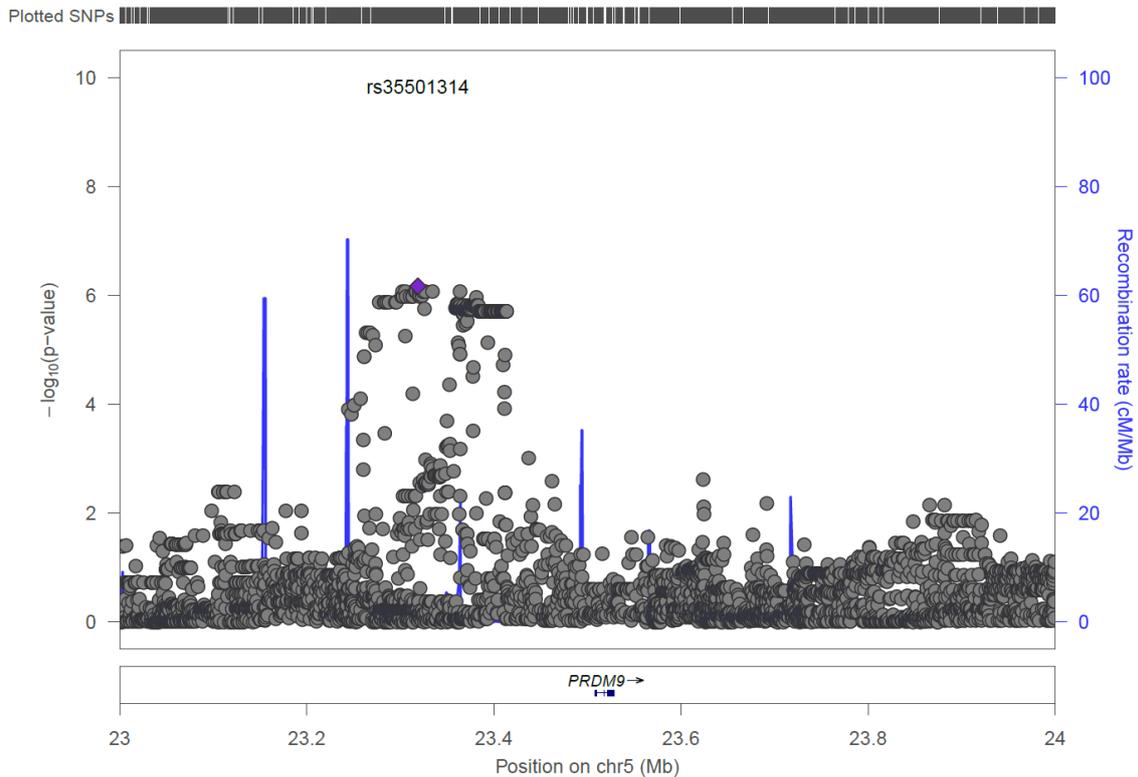
**Figure 74** – Regional association plot showing EBV copy number association with rs11668582 and the LTPB4 locus.

In addition, association was seen for a genomic region on chromosome 5 with 15 SNPs associated with EBV copy number (Figure 75-76).



**Figure 75** – Plot showing chromosome 5 and EBV copy number associations, with the PRDM9 association peak; negative log of the P-value indicated on the X-axis

The SNPs are upstream of *PRDM9*, a gene that encodes a zinc-finger DNA-binding protein expressed by germline cells and controlling meiotic recombination (Myers et al. 2010, Baudat et al.2010).



**Figure 76** – A regional association plot showing SNPs in the PRDM9 locus and EBV copy number associations.

The results are summarised in Table 13, below.

SNP	Chr	P-value	gene	left_gene	right_gene	eQTL for
rs192833696	1	3.88E-07				
rs77407282	2	7.49E-07				
rs17379786	2	2.47E-06	EPHA4	LOC100129746	LOC729770	
rs35501314	5	6.82E-07		LOC391771	PRDM9	
rs77620363	5	8.51E-07				
rs201034572	5	8.51E-07				
rs34549526	5	8.51E-07		LOC391771	PRDM9	
rs34075772	5	8.51E-07		LOC391771	PRDM9	
rs34803213	5	8.51E-07		LOC391771	PRDM9	
rs71611225	5	8.51E-07				
rs75726391	5	8.51E-07				
rs71611228	5	8.51E-07				
rs6861201	5	8.51E-07		LOC391771	PRDM9	EMR2; DJC12; RBX1
rs13180619	5	8.51E-07		LOC391771	PRDM9	
rs71611230	5	8.51E-07				
rs59556584	5	8.52E-07		LOC391771	PRDM9	
rs71611226	5	9.52E-07				
rs35624884	5	9.98E-07		LOC391771	PRDM9	
rs1609819	8	6.60E-07		PRAGMIN	CLDN23	
rs143723136	11	8.49E-07				
rs185368921	11	8.91E-07				
<b>rs190636588</b>	<b>14</b>	<b>2.16E-08</b>				
rs10873551	14	7.87E-07		RPS2P4	KIAA0284	
rs4513132	17	5.86E-07		KCTD2	SLC16A5	
rs4789129	17	6.01E-07		KCTD2	SLC16A5	
rs11668582	19	6.11E-07	LTBP4	SHKBP1	NUMBL	
rs12980111	19	6.95E-07	LTBP4	SHKBP1	NUMBL	
rs11667706	19	7.90E-07	LTBP4	SHKBP1	NUMBL	
rs8107014	19	8.42E-07	LTBP4	SHKBP1	NUMBL	

**Table 13** - EBV copy number QTLs from the 1000 Genomes samples:

- 1) “SNP” - significant and suggestive ( $P\text{-value} < 1E^{-06}$ ) candidates (SNPs) are listed in the first column,
- 2) “Chr” – genomic location of candidate SNPs is given by the second column,
- 3) “P-value” P-values for the relevant SNP’s test statistics are indicated by the third field,
- 4) “gene” – indicates the name of the gene a candidate QTL is located in,
- 5) “left” and 6) “right gene” - name the two closest flanking genomic locations,
- 7) “eQTL for” lists the human gene for which the SNP acts as a significant eQTL according to one of the eQTL databases – SCAN, Genevar and GTEx.

## 4.5 Discussion

### 4.5.1 Copy number QTLs in HapMap LCLs

Association tests carried out using copy number data from the HapMap cell lines gave no significant results. In fact, only a single association peak was observed within the suggestive range in the Yoruban LCLs. The peak was located in *CDH12*, and the SNP with the lowest P-value was rs4489023. *CDH12* belongs to the cadherin superfamily and is thought to be involved in calcium-dependent cell adhesion or formation and maintenance of synaptic structures (Zhao 2013; Hansell et al. 2010). Its functions have been poorly characterised, and recently has been reported as a potential tumour suppressor responsible for lung cancer progression and susceptible to age-dependent methylation (Zhao 2013, Damaschke et al. 2013). The candidate has no proven or immediate link to tumours of the lymphatic tissue or Herpesvirus infection.

### 4.5.2 Copy number QTLs in 1000 Genomes samples

A number of associations with EBV copy number were identified although only one achieved genome-wide significance (rs190636588 on chromosome 14). Biologically, at least two other striking associations were found.

One involved 4 SNPs within *LTBP4* gene. *LTBP4* is responsible for secretion of the TGF- $\beta$ , a pluripotent, immunosuppressive anti-inflammatory cytokine (Hawinkels and Ten-Dijke 2010) which may favour tumour survival and growth. Many human tumour cells secrete TGF that hampers antigen-specific cellular immunity (Foster et al. 2008). For example, TGF- $\beta$  derived from Epstein-Barr virus HL and non-Hodgkin lymphoma tumour cells decreases the

efficiency of EBV-specific cytotoxic T cells in eliminating them (Foster et al. 2008). HL cells express latency type II transcripts and present EBV antigens however they also produce TGF- $\beta$  which inhibits T-cell immunity (Foster et al. 2008). TGF- $\beta$  also inhibits growth in normal [human](#) B-cells and BL cells (Chaouchi et al. 1995). B-cells have high-affinity receptors for TGF- $\beta$ , and at the same time, they secrete endogenous TGF- $\beta$ , which might be an autocrine mechanism that controls their growth and differentiation (Kehrl et al. 1986). On the other hand LCLs do not secrete endogenous TGF- $\beta$  and have a decreased level of expression of its receptors (Willet et al. 2013, Wroblewski et al. 2002).

TGF- $\beta$  secreted by T cells can induce EBV reactivation mediated by *BZLF1* gene (Kenney and Mertz 2014). There is evidence that activation of endogenous TGF- $\beta$  production is capable of disrupting EBV latency and inducing the lytic replication, coming from BL cell studies (di Renzo et al. 1994). When TGF- $\beta$  binds its cognate receptor it activates SMAD and MAPK pathways, which is essential for EBV reactivation (Murata et al. 2014 2013). Together with other cytokines, for example, viral IL-10 (vIL-10 - encoded by EBV) it is produced during the lytic cycle. These cellular cytokines are secreted from cells replicating lytically and act to propagate lytic replication in the nearest B-cells in which EBV is latent (Murata and Tsurumi 2013). In a recent analysis it was found that BCR signalling but also TGF- $\beta$  stimulation and hypoxia are a major critical cause of EBV reactivation (Murata and Tsurumi 2013). Upon initiation of lytic replication viral protein, BZLF1, induces secretion of endogenous TGF- $\beta$ , establishing a positive feedback loop that promotes further lytic replication and immuno-suppression (Murata and Tsurumi 2013). TGF- $\beta$  is upregulated and can act synergistically with LMP1 to induce expression of certain genes (Sides et al 2011). Therefore in addition to direct stimulation of lytic EBV replication by both endo- and exogenous application, TGF- $\beta$  stimulation might also play a role at latency II germinal centre lymphoblast rescue stage. Particularly that both BCR engagement, substituted mainly by

LMP2, and TGF- $\beta$  signalling contribute to NF- $\kappa$ B activation (Kenney and Mertz 2014). Consequently, there is evidence linking TGF stimulation directly to the ability of EBV to propagate itself lytically and re-infect other B-cells or, alternatively, establish Latency II. TGF could, together with other cytokines and factors, act as a switch determining whether the virus leaves the cell or remains silent. This could potentially contribute to copy number through increased number of lytically replicating and re-infected B-cells. LTBP4 could attenuate the effects of TGF- $\beta$  as it binds TGF- $\beta$  upon its transportation and release into extracellular matrix, where LTBP4-bound TGF- $\beta$  remains inactive and sheds LTBP4 when activated (NCBI website, Gene ID: 8425). *LTBP4* mutations can have a range of pleiotropic effects and result in physiological abnormalities (NCBI website, Gene ID: 8425).

A second potentially very interesting association peak linked EBV copy number to SNPs within *PRDM9*. This highly polymorphic zinc finger protein determines the sites of homologous recombination during meiosis. *PRDM9* binds to DNA and using its SET domain to locally tri-methylate histone H3 at lysine 4, targeting the site for double strand break and Holliday junction formation as well as genetic recombination (Baker et al. 2014). The diversity of *PRDM9* repetitive “zinc finger” motifs could influence genomic instability and certain rare variants of *PRDM9* could increase the chances of genomic rearrangements associated with childhood leukaemias (Hussin et al. 2013). However, no link to viral infection or lymphoproliferative disease has been reported.

In addition, according to SCAN database, one of these copy number associated variants has been identified as an eQTL for Ring-box protein 1, *RBX1*, part of VHL tumour suppressor complex and SCF ubiquitin ligase. *Rbx1* has been reported to regulate LMP1-induced NF- $\kappa$ B signalling by proteasome-dependent degradation of the inhibitor of kappa-B (*I $\kappa$ B $\alpha$* ), which reduces inhibition of NF- $\kappa$ B-dependent transcription (Kanarek and Ben-Neriah 2012). Another study implicated *RBX1* in promoting HIF1 $\alpha$  proteasomal degradation. The same

study reported that LMP1 increases HIF1 $\alpha$  retention (Kondo et al. 2002). Suzuki et al. (2010) reported that latency-associated nuclear antigen (LANA) of KSHV, a virus most related EBV, interacts with RBX1 and promotes ubiquitylation and degradation of tumour suppressors, in particular p53.

Finally, association with copy number was noted for two SNPs located within and proximal to the gene encoding Eph Receptor 4A (*EPHA4*), however no association peak was present. Epha4 belongs to the family of Eph receptor tyrosine kinases, transmembrane proteins which interact with A and B-type ephrins and play a role in developing nervous system's axonal repulsion and synapse formation, synaptic plasticity and memory (Van Hoecke et al. 2012). EphA4 is expressed in multiple tissues and Eph receptors also play a role in controlling physiology and homeostasis in adult organs and their malfunction could lead to disease (Swartz et al. 2001, Pasquale 2008, Munro et al. 2012&2013). A range of other cellular functions of Eph including actin cytoskeleton formation, cell adhesion, shape, movement, proliferation, survival, differentiation, and secretion have been proposed (Pasquale 2008, Munoz et al. 2002). It is interesting to note that recently EphA4 has been reported to modulate the onset and the course of experimental autoimmune encephalomyelitis EAE, which is an animal model of MS, with EphA4 knock-outs exhibiting less severe symptoms and reduced axonal pathology (Munro et al. 2013). What is more, EphA4 has been reported to be the cellular receptor used by Kaposi sarcoma herpesvirus (KSHV, or human herpesvirus 8 - HHV8), highly related to EBV, as well as rhesus monkey rhadinovirus (RRV) to enter endothelial and B-cells (Hahn and Desrosiers 2013, Hahn et al. 2012). There is some evidence that *EPHA4* may play a role in lytic switch of KSHV, although Tousel-like kinases (TLKs) grossly supersede this effect (Dillon et al. 2013).

A clear although not statistically significant association peak was present on chromosome 17. This peak was located within an intergenic region, approximately halfway in between the

potassium channel tetramerization domain containing 2 (*KCTD2*) ORF and solute carrier family 16 (monocarboxylate transporter) member 5 (*SLC16A5*) gene – both playing important roles in cellular metabolism but with no apparent link to EBV or B-cell biology.

### 4.5.3 Conclusion

Overall, the population-based approach, larger samples size and denser genotyping contributed to the copy number QTL mapping giving more robust associations than the latency eQTL testing described in chapter 3. Three distinct association peaks of biological interest for further study were identified, notably involving *LTBP4* and *PRMD9* in addition to two associations close to *EPHA4*. Particularly *LTBP4* appears to be directly involved in EBV transformation through its role in TGF- $\beta$  secretion. If mutations in this gene are indeed affecting the rate of TGF- $\beta$  secretion and activation, then *LTBP4* may be directly contributing to the copy number of EBV. This is because TGF- $\beta$ , which normally appears to keep proliferation of B-cells in check by endogenous autocrine regulation, elicits lytic reactivation of EBV in BL-cells characterised by Latency I or, less typically II (Letterio et al. 1998, di Renzo et al. 1994). Thus, although LCLs expressing Latency III transcripts are largely insensitive to its signalling (Willet et al. 2013), TGF- $\beta$  could still prove important for the B-cell survival *in vivo*, particularly at the germinal stage characterised Latency II. It is also interesting to note that three members of the cadherin family appeared as candidates in both latency eQTL and copy number assays (Choy et al. 2008 data).

## **Chapter 5. Meta-analysis**

### **5.1. Introduction**

EBV latency and copy number QTL assays were carried out in multiple independent cohorts, most of which had a relatively small sample size of less than 100 individuals. This is likely to account for significant associations not having been replicated across any two cohorts. To integrate the findings from all the experiments and check the overall consistency and direction of the associations, it was necessary to apply a testing method for the combined results. This did should not just improve the power to detect associations by increasing the effective sample size, but also make it possible to sort them by giving highest priority to those which are consistently replicated across all assayed cohorts with the same direction of effect on gene expression. In addition, relationships between copy number, EBV expression and human expression could be explored by conducting correlation tests.

### **5.2 Aims of the chapter**

1. To perform a meta-analysis for the overlapping phenotypes in all available cohorts in order to integrate and prioritise the findings.
2. To integrate the new meta-analysed EBV results with MRC-A human eQTLs as well as GWAS Catalog loci.
3. To correlate EBV latency transcript expression with human expression in the MRC-A samples.
4. To correlate EBV transcript expression levels with copy number for those LCLs for which both are available.

## 5.3 Results

### 5.3.1 Latency eQTL meta-analysis

Meta-analysis of EBV latency eQTLs combined from three cohorts (MRC-A, and two RNA-seq study populations - CEPH as well as Yoruba) provided multiple significant and suggestive associations for all 6 transcripts tested.

#### **Quality control**

Before meta-analysis was carried out, to ensure better uniformity, Z-score normalisation was applied to RNA-seq EBV expression data quantified by Arvey et al (2011) (MRCA panel expression data had already been subjected to Z-score normalisation).

#### **Results**

Table 15 summarises the 55 top associations, with P-values below  $1E^{-06}$ . Manhattan and QQ plots are also presented (Figures 77, 81-84, 87). As expected, some heterogeneity of effects was detected, however the small number of cohorts and samples meant for most of the top associations it is not significant. For most of the SNPs genotyped in all of the tested cohorts, the heterogeneity indicated by  $I^2$  index was none or low (>25%), with two SNPs (rs10492785 and rs6825926) exhibiting high level (~ or >75%) as well as statistical significance (P-value < 0.05) for the corresponding Cochran's Q statistic. For all but three of the SNPs that showed a heterogeneous effect on expression, an opposite direction of the effect was present in one cohort versus the others (including rs10492785, rs6825926). Overall, the results indicate moderate heterogeneity present, which suggests a random effect metanalysis (which assumes

that the SNP effects are not identical across all studies, but should follow a distribution) could be informative. However, because the experiment includes two relatively small panels and a larger one (MRCA), heterogeneity test with random effects meta-analysis may be skewed by the smaller cohorts resulting in wider confidence intervals for the random effects. A fixed-effect analysis gives more weight to the effect of the single larger cohort. Since the MRCA cohort was the primary source of information on the latency eQTLs in the current study, as well as subjected to a thorough qPCR assay in the lab, a fixed effect meta-analysis was chosen.

The table below (Table 14) reports the values of the genomic control inflation factor, lambda. By reporting the value of the lambda factor, GWAMA provides means to control for population structure and correct the test statistics, if necessary. Lambda is the median of the test statistics divided by its expectation under the null hypothesis of no association (Magi and Morris 2010).

**Table 14** – Values of the lambda genomic control factor for tested populations.

Gene	Cohort	Lambda
EBER1	MRCA	1.13
	Arvey CEPH	1.02
	Arvey YRI	1.03
EBER2	MRCA	0.92
	Arvey CEPH	1.02
	Arvey YRI	1.04
EBNA1	MRCA	1.06
	Arvey CEPH	1.02
	Arvey YRI	0.98
EBNA2	MRCA	1.18
	Arvey CEPH	0.99
	Arvey YRI	1.01
EBNA3A	MRCA	1.005
	Arvey CEPH	1.073
	Arvey YRI	0.963
LMP1	MRCA	1.08
	Arvey CEPH	0.98
	Arvey YRI	1.01

It is thus the factor by which the association test statistics are increased due to population structure and by which they ought to be uniformly multiplied to correct for the inflation of P-values (Devlin and Roeder 1999). Among the populations tested in the current meta-analysis, the genomic control parameter exceeded the recommended range of 0.95-1.05 in the MRC-A samples (particularly when tested for EBER1 and EBNA2 associations). This indicates an inflation of P-values, especially in the EBER1 EBNA2 association tests. Genomic control should be applied to account for population stratification and reduce the lambda value (Devlin and Roeder 1999). The `-gc` option (to output the genomic control lambda factor values in a separate `gc.out` file) and the `-gco` (to correct for the inflation where necessary) were specified. Statistically significant associations are shown in bold and solitary associations against the white background. The first column (“SNP”) lists the rs number of the associated SNP and the 4<sup>th</sup> contains corresponding P-values. Chromosome and base pair position is given by the 2<sup>nd</sup> and 3<sup>rd</sup> column. Column 5 “Studies” informs how many cohorts had a genotype available for a given SNP and column 6 “Samples” lists the number of samples available for meta-analysis. Next column, “Effect” shows the direction of the SNP’s regulatory effect on transcript levels in each cohort (“-“ if minor allele dosage is inversely correlated with transcript abundance). Column 8, “Trait” lists the tested phenotype and the final three columns indicate whether the variant is located within a gene and provide its name (“Gene” column) as well as list the two closest flanking genes.

SNP	Chr	BP	p-value	Studies	Samples	Effect	Trait	Gene	Left Gene	Right Gene
rs6825926	4	122939162	5.01E-07	3	408	++	EBER1	NA	TMEM155	EXOSC9
<b>rs6836544</b>	<b>4</b>	<b>122945523</b>	<b>8.84E-09</b>	<b>3</b>	<b>409</b>	<b>++</b>	<b>EBER1</b>	<b>EXOSC9</b>	<b>TMEM155</b>	<b>CCNA2</b>
rs769243	4	122961257	9.00E-08	3	409	++	EBER1	CCNA2	EXOSC9	BBS7
<b>rs1017696</b>	<b>8</b>	<b>2372758</b>	<b>1.06E-08</b>	<b>2</b>	<b>128</b>	<b>?--</b>	<b>EBER1</b>	<b>NA</b>	<b>MYOM2</b>	<b>CSMD1</b>
rs10761703	10	64532968	7.87E-07	3	409	---	EBER1	NA	EGR2	NRBF2
rs7994857	13	95868045	9.37E-07	3	408	---	EBER1	HS6ST3	UGCGL2	HSP90AB6P
rs7153369	14	61601967	4.26E-07	3	409	---	EBER1	SYT16	MOC53P	LOC100130953
rs7161387	14	94591164	5.13E-07	3	409	++	EBER1	NA	LOC730118	DICER1
rs10134473	14	94592888	8.54E-07	3	409	++	EBER1	NA	LOC730118	DICER1
rs12918328	16	6752939	1.77E-07	3	409	++	EBER1	A2BP1	LOC100131413	LOC100131080
rs7191315	16	6753913	1.18E-07	3	409	++	EBER1	A2BP1	LOC100131413	LOC100131080
rs11867159	16	6756183	4.36E-07	1	277	++?	EBER1	A2BP1	LOC100131413	LOC100131080
rs2733518	3	17224607	3.51E-07	3	409	---	EBER2	TBC1D5	PLCL2	TBC1D5
<b>rs12495053</b>	<b>3</b>	<b>24375132</b>	<b>3.30E-12</b>	<b>3</b>	<b>409</b>	<b>+++</b>	<b>EBER2</b>	<b>THRB</b>	<b>NR1D2</b>	<b>LOC644990</b>
rs7764708	6	124243625	5.12E-07	2	132	?++	EBER2	NKAIN2	TRDN	RNF217
rs12277302	11	14082329	9.65E-07	3	409	---	EBER2	SPON1	LOC729147	RRAS2
<b>rs794152</b>	<b>12</b>	<b>30687458</b>	<b>4.53E-08</b>	<b>2</b>	<b>342</b>	<b>--?</b>	<b>EBER2</b>	<b>IPO8</b>	<b>TMTC1</b>	<b>CAPRIN2</b>
rs2645924	13	55348356	6.59E-07	1	67	?+?	EBER2	NA	LOC100128485	LOC100130462
rs10492785	16	19454858	4.91E-07	3	409	++	EBER2	CP110	GDE1	C16orf62
rs41423044	1	110647437	5.90E-07	1	65	?+?	EBNA1	NA	KCNC4	RBM15
rs1486184	2	108067002	1.99E-07	3	408	---	EBNA1	NA	SLC5A7	SULT1C3
rs1290443	3	25576839	7.39E-07	3	409	+++	EBNA1	RARB	LOC100130354	TOP2B
<b>rs17158630</b>	<b>10</b>	<b>43046865</b>	<b>3.02E-08</b>	<b>1</b>	<b>277</b>	<b>-??</b>	<b>EBNA1</b>	<b>RASGEF1A</b>	<b>CSGALNACT2</b>	<b>FXYD4</b>
rs367312	13	113948297	2.30E-07	3	409	---	EBNA1	NA	RASA3	CDC16
rs9613538	22	26576308	8.47E-08	2	132	?+?	EBNA1	PITPNB	MN1	PITPNB
rs10739626	9	125113651	3.78E-07	3	407	---	EBNA2	NA	STRBP	CRB2
rs6482682	10	130187972	7.77E-07	3	409	---	EBNA2	NA	MKI67	hCG_1795091
rs12411264	1	170460291	4.00E-07	3	409	++	EBNA3A	LOC100128178	LOC127099	LOC100131486
rs12410416	1	170460443	7.58E-07	3	409	++	EBNA3A	DNM3	LOC127099	LOC100131486
rs2289635	1	170475243	6.49E-07	2	132	?+-	EBNA3A	DNM3	LOC127099	LOC100131486
rs4072117	1	170475494	3.48E-07	3	409	++	EBNA3A	LOC100128178	LOC127099	LOC100131486
rs6678749	1	170484406	3.44E-07	3	409	++	EBNA3A	LOC100128178	LOC127099	LOC100131486
rs12075079	1	170486618	3.38E-07	3	409	++	EBNA3A	LOC100128178	LOC127099	LOC100131486
rs16844255	1	170537673	1.11E-07	2	342	++?	EBNA3A	DNM3	LOC100128178	LOC100131486
rs949571	1	195579812	7.91E-07	3	409	+++	EBNA3A	CRB1	LOC127011	MRPS21P3
rs7660783	4	8727943	7.24E-07	3	408	+++	EBNA3A	NA	CPZ	LOC728369
rs731129	5	172410515	7.33E-07	1	67	?-?	EBNA3A	NA	ATP6VOE1	C5orf41
rs4959235	6	3304336	1.29E-07	2	342	++?	EBNA3A	SLC22A23	PSMG4	LOC643327
rs17060278	6	131814081	3.45E-07	1	66	?-?	EBNA3A	NA	AKAP7	ARG1
rs9346605	6	169159366	6.18E-07	3	409	+++	EBNA3A	NA	SMOC2	LOC100132959
rs11812728	10	130858043	7.33E-07	1	67	?-?	EBNA3A	NA	hCG_1795091	MGMT
<b>rs7162181</b>	<b>15</b>	<b>78606479</b>	<b>9.23E-07</b>	<b>3</b>	<b>408</b>	<b>+++</b>	<b>EBNA3A</b>	<b>ARNT2</b>	<b>LOC730126</b>	<b>FAM108C1</b>
<b>rs11856676</b>	<b>15</b>	<b>78609602</b>	<b>5.58E-07</b>	<b>3</b>	<b>408</b>	<b>+++</b>	<b>EBNA3A</b>	<b>ARNT2</b>	<b>LOC730126</b>	<b>FAM108C1</b>
<b>rs4238523</b>	<b>15</b>	<b>78616704</b>	<b>6.83E-07</b>	<b>3</b>	<b>409</b>	<b>+++</b>	<b>EBNA3A</b>	<b>ARNT2</b>	<b>LOC730126</b>	<b>FAM108C1</b>
rs9935655	16	20250398	1.57E-07	3	409	---	EBNA3A	UMOD	GP2	UMOD
rs13347021	19	6411264	7.33E-07	1	67	?-?	EBNA3A	SLC25A23	SLC25A23	CRB3
rs8107408	19	36790326	1.09E-07	1	67	?-?	EBNA3A	NA	TSHZ3	ZNF507
rs16856218	2	131348278	7.23E-07	1	277	++?	LMP1	LOC730032	CYCSP8	ARHGEF4
rs3796349	3	53808549	8.48E-07	3	409	+++	LMP1	CACNA1D	LOC391539	CHDH
rs10899353	11	76515200	4.10E-07	3	409	+++	LMP1	CAPN5	CAPN5	MYO7A
rs2035319	12	64596813	3.72E-07	1	277	++?	LMP1	HMG2A	LOC100129940	LOC729484
rs12322842	12	96094441	9.97E-07	1	277	++?	LMP1	NA	NEDD1	RMST
rs4899192	14	65569935	9.32E-07	1	277	++?	LMP1	NA	YBX1P1	C14orf53
rs6499709	16	53247141	9.78E-07	3	408	---	LMP1	NA	LOC728792	hCG_1815491
rs4815837	20	5473648	4.61E-07	3	409	+++	LMP1	RP5-1022P6.2	RPS18P1	EIF4EP1

**Table 15** – Meta analysis of MRC-A and RNA-seq latency eQTLs. Legend in the main text.

rs number	q statistic	q p-value	I <sup>2</sup>	n studies	n samples
rs6825926	7.07	0.03	<b>0.72</b>	3	408
<b>rs6836544</b>	<b>3.18</b>	<b>0.20</b>	<b>0.37</b>	<b>3</b>	<b>409</b>
rs769243	3.06	0.22	<b>0.35</b>	3	409
<b>rs1017696</b>	<b>0.02</b>	<b>0.87</b>	<b>0.00</b>	<b>2</b>	<b>128</b>
rs10761703	1.31	0.52	0.00	3	409
rs7994857	1.45	0.49	0.00	3	408
rs7153369	0.44	0.80	0.00	3	409
rs7161387	1.60	0.45	0.00	3	409
rs10134473	0.93	0.63	0.00	3	409
rs12918328	4.84	0.09	<b>0.59</b>	3	409
rs7191315	4.63	0.10	<b>0.57</b>	3	409
rs11867159	0.00	1.00	NA	1	277
rs2733518	5.34	0.07	<b>0.63</b>	3	409
<b>rs12495053</b>	<b>0.45</b>	<b>0.80</b>	<b>0.00</b>	<b>3</b>	<b>409</b>
rs7764708	0.90	0.34	0.00	2	132
rs12277302	2.16	0.34	<b>0.08</b>	3	409
<b>rs794152</b>	<b>0.84</b>	<b>0.36</b>	<b>0.00</b>	<b>2</b>	<b>342</b>
rs2645924	0.00	1.00	NA	1	67
rs10492785	13.27	0.0013	<b>0.85</b>	3	409
rs41423044	0.00	1.00	NA	1	65
rs1486184	4.88	0.09	<b>0.59</b>	3	408
rs1290443	2.26	0.32	0.12	3	409
<b>rs17158630</b>	<b>0.00</b>	<b>1.00</b>	<b>NA</b>	<b>1</b>	<b>277</b>
rs367312	0.11	0.95	0.00	3	409
rs9613538	1.75	0.19	<b>0.43</b>	2	132
rs10739626	2.09	0.35	0.04	3	407
rs6482682	0.33	0.85	0.00	3	409
rs12411264	2.63	0.27	<b>0.24</b>	3	409
rs12410416	4.11	0.13	<b>0.51</b>	3	409
rs2289635	2.29	0.13	<b>0.56</b>	2	132
rs4072117	2.56	0.28	<b>0.22</b>	3	409
rs6678749	2.55	0.28	<b>0.22</b>	3	409
rs12075079	2.50	0.29	<b>0.20</b>	3	409
rs16844255	0.23	0.63	0.00	2	342
rs949571	0.03	0.99	0.00	3	409
rs7660783	0.63	0.73	0.00	3	408
rs731129	0.00	1.00	NA	1	67
rs4959235	0.98	0.32	0.00	2	342
rs17060278	0.00	1.00	NA	1	66
rs9346605	6.18	0.05	<b>0.68</b>	3	409
rs11812728	0.00	1.00	NA	1	67
rs7162181	0.27	0.87	0.00	3	408
rs11856676	0.21	0.90	0.00	3	408
rs4238523	0.20	0.91	0.00	3	409
rs9935655	1.71	0.43	0.00	3	409
rs13347021	0.00	1.00	NA	1	67
rs8107408	0.00	1.00	NA	1	67
rs16856218	0.00	1.00	NA	1	277
rs3796349	0.70	0.70	0.00	3	409
rs10899353	0.12	0.94	0.00	3	409
rs2035319	0.00	1.00	NA	1	277
rs12322842	0.00	1.00	NA	1	277
rs4899192	0.00	1.00	NA	1	277
rs6499709	0.11	0.95	0.00	3	408
rs4815837	0.49	0.78	0.00	3	409

**Table 15A** – Meta analysis of MRC-A and RNA-seq latency eQTLs. Table presents the extent of effect heterogeneity across tested cohorts for all the SNPs in Table 15. Cochran’s Q statistic and its corresponding P-value are given in the 2<sup>nd</sup> and 3<sup>rd</sup> columns, while the I<sup>2</sup> index (proportion of total variability explained by heterogeneity) in the 4<sup>th</sup>.

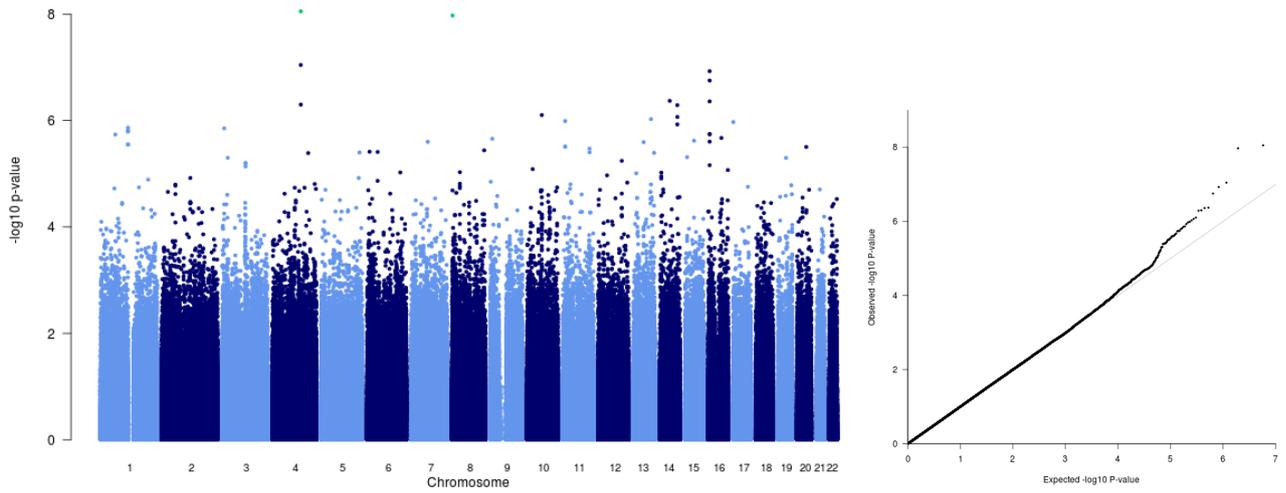
Most of the identified candidate eQTLs fall within the suggestive range and consist of single SNPs. There are only five loci for which the association involves more than a single SNP. Three such associations involved EBV1 including a region containing *CCNA2* and *EXOSC9* on chromosome 4 (Figure 78), a locus located ~30kb upstream from the 3' end of *DICER1* on chromosome 14 (Figure 79), and *RBFOX1* (also known as *A2BPI*) (Figure 80). The other two associations were for EBNA3A transcript levels, localising within *ARNT2* on chromosome 15 (Figure 85) and within *DNM3* on chromosome one (Figure 86). None of the results included in Table 15 had been described in previous analysis, mainly due to non-significance, with the exception of rs17158630, located within *RASGEF1A*, and rs11867159, located within *RBFOX1*. Both are present in the MRC-A latency eQTL results. However, rs17158630 was one of 13 SNPs whose genotypes were known only for a single cohort

Overlap with the NHGRI GWAS Catalog was investigated (Table 16). This indicated that one of the associated loci on chromosome 1, partly overlapping with *DNM3*, contains loci associated with height and bone mineral density. When viral vs human eQTL overlap was investigated, a solitary association on chromosome 6 within *SLC22A23* was linked to antipsychotic-induced QTc interval (an interval in human heart's electrical cycle) prolongation. A single SNP, rs367312, was also identified as an eQTL in the MRC-A human expression results (FDR 0.003). The associated transcript belonged to *UPF3A*, a nuclear protein and component of mRNA postsplicing complex, with no links to EBV.

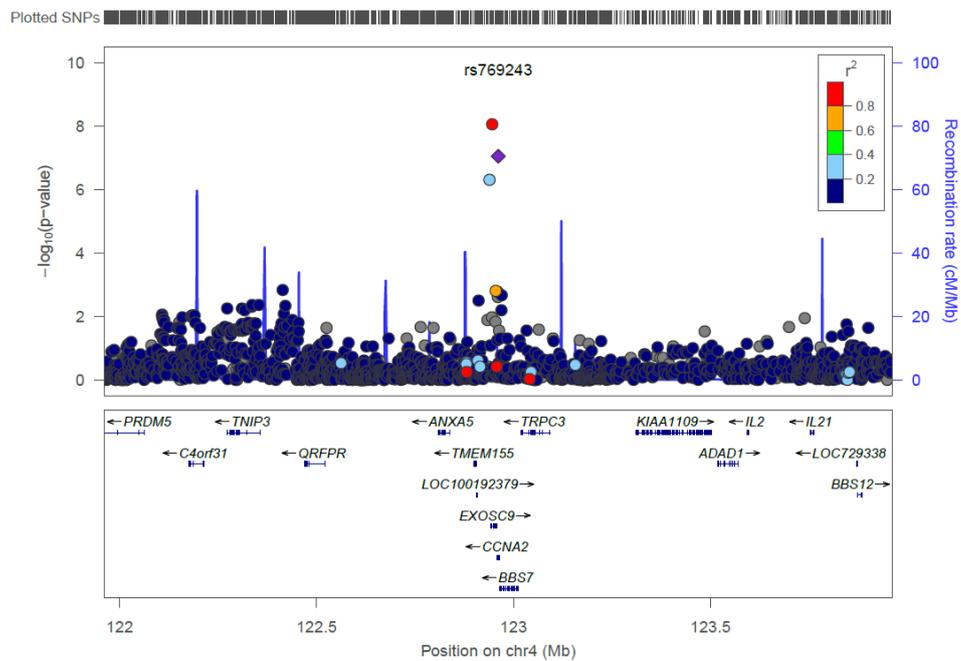
SNP	Chr	p-value	Studies	Samples	Effect	Trait	Gene	GWAS.Disease.Trait	GWAS.Gene	p-value
rs12411264	1	4.00E-07	3	409	++-	EBNA3A	LOC10012	Height	DNM3	7.00E-12
rs12411264	1	4.00E-07	3	409	++-	EBNA3A	LOC10012	Height	DNM3	3.00E-08
rs12411264	1	4.00E-07	3	409	++-	EBNA3A	LOC10012	Bonemineraldensity	DNM3	9.00E-15
rs12410416	1	7.58E-07	3	409	++-	EBNA3A	DNM3	Height	DNM3	3.00E-08
rs12410416	1	7.58E-07	3	409	++-	EBNA3A	DNM3	Height	DNM3	7.00E-12
rs12410416	1	7.58E-07	3	409	++-	EBNA3A	DNM3	Bonemineraldensity	DNM3	9.00E-15
rs2289635	1	6.49E-07	2	132	?+-	EBNA3A	DNM3	Height	DNM3	3.00E-08
rs2289635	1	6.49E-07	2	132	?+-	EBNA3A	DNM3	Bonemineraldensity	DNM3	9.00E-15
rs2289635	1	6.49E-07	2	132	?+-	EBNA3A	DNM3	Height	DNM3	7.00E-12
rs4072117	1	3.48E-07	3	409	++-	EBNA3A	LOC10012	Height	DNM3	9.00E-15
rs4072117	1	3.48E-07	3	409	++-	EBNA3A	LOC10012	Height	DNM3	3.00E-08
rs4072117	1	3.48E-07	3	409	++-	EBNA3A	LOC10012	Height	DNM3	7.00E-12
rs6678749	1	3.44E-07	3	409	++-	EBNA3A	LOC10012	Height	DNM3	3.00E-08
rs6678749	1	3.44E-07	3	409	++-	EBNA3A	LOC10012	Bonemineraldensity	DNM3	9.00E-15
rs6678749	1	3.44E-07	3	409	++-	EBNA3A	LOC10012	Height	DNM3	7.00E-12
rs12075079	1	3.38E-07	3	409	++-	EBNA3A	LOC10012	Height	DNM3	7.00E-12
rs12075079	1	3.38E-07	3	409	++-	EBNA3A	LOC10012	Height	DNM3	3.00E-08
rs12075079	1	3.38E-07	3	409	++-	EBNA3A	LOC10012	Bonemineraldensity	DNM3	9.00E-15
rs16844255	1	1.11E-07	2	342	++?	EBNA3A	DNM3	Height	DNM3	3.00E-08
rs16844255	1	1.11E-07	2	342	++?	EBNA3A	DNM3	Bonemineraldensity	DNM3	9.00E-15
rs16844255	1	1.11E-07	2	342	++?	EBNA3A	DNM3	Height	DNM3	7.00E-12
rs4959235	6	1.29E-07	2	342	++?	EBNA3A	SLC22A23	Antipsychotic	SLC22A23	2.00E-07

**Table 16** – Overlap between suggestive latency eQTLs from the MRC-A/RNAseq data meta-analysis, and GWAS Catalog risk loci. Legend – fields 1 to 8 as in **Table 15**. Field 9 “GWAS.Disease.Trait.” indicates the disease or trait for which a significant association was found to a relevant SNP. Field 10 “GWAS.Gene” reports the gene proposed to be regulated by the SNP associated with the disease. The final field “P-value” lists the P-value of the GWAS test statistic for the association between the SNP and the disease/trait.

An overview of the meta-analysis association results for all 6 tested transcripts is given by the Manhattan plots below. Each genome-wide association plot is accompanied by a corresponding QQ plot indicating the deviation of the test statistics from the expected distribution under the null hypothesis. For chosen associations from Table 15, regional association plots constructed using LocusZoom are also shown. It is important to note that regional plots were constructed with the “Genome Build/LD Population” setting specified as “hg19/1000 Genomes March 2012 EUR”. This means only the European 1000 Genomes populations (Finnish in Finland, Utah Residents (CEPH) with Northern and Western European ancestry, Toscani in Italy, British in England and Scotland, Iberian population in Spain) were used to estimate the LD structure, however the current study included also African Yoruba samples. As a result, some of the associations presented by the regional plots follow an unusual, irregular LD pattern.

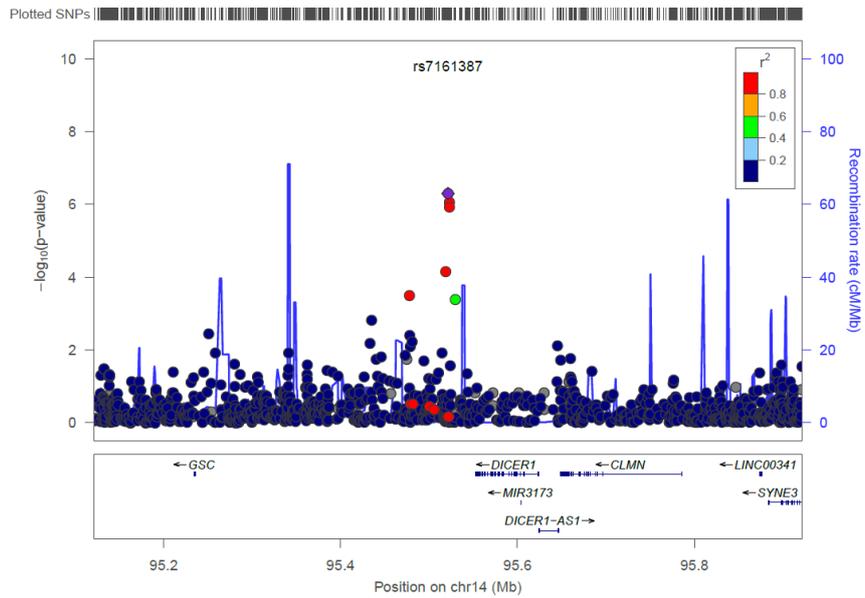


**Figure 77** - EBER 1 genome-wide association plot showing results of MRC-A/RNA-seq meta-analysis and corresponding QQ plot; negative log of P-value given by the Y axis; chromosome and expected negative of log P-value shown on the X-axis.



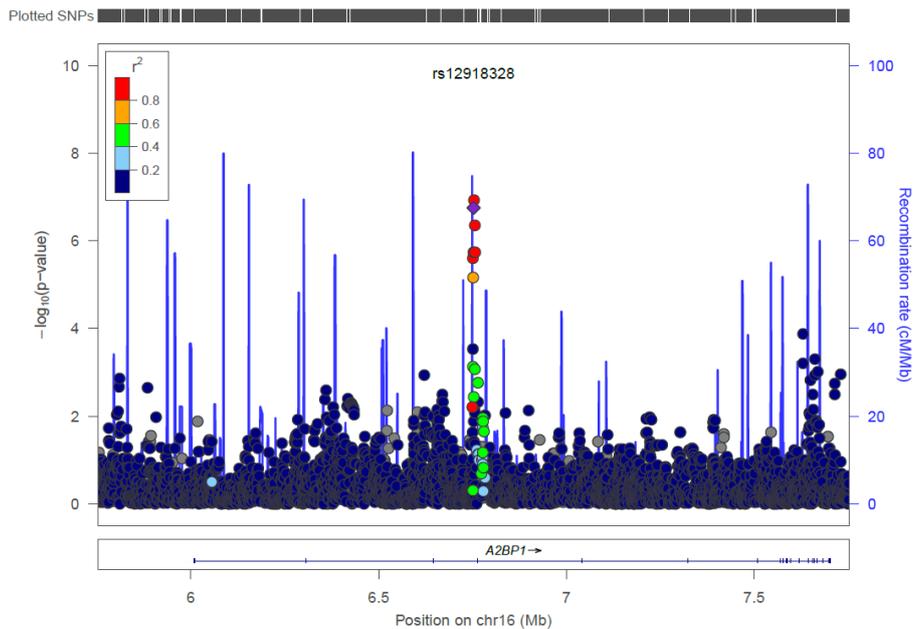
Note: LD structure is based on the 1000 Genomes European reference data.

**Figure 78** - Regional association plot showing association between EBER1 RNA levels and rs4513132 in CCNA2



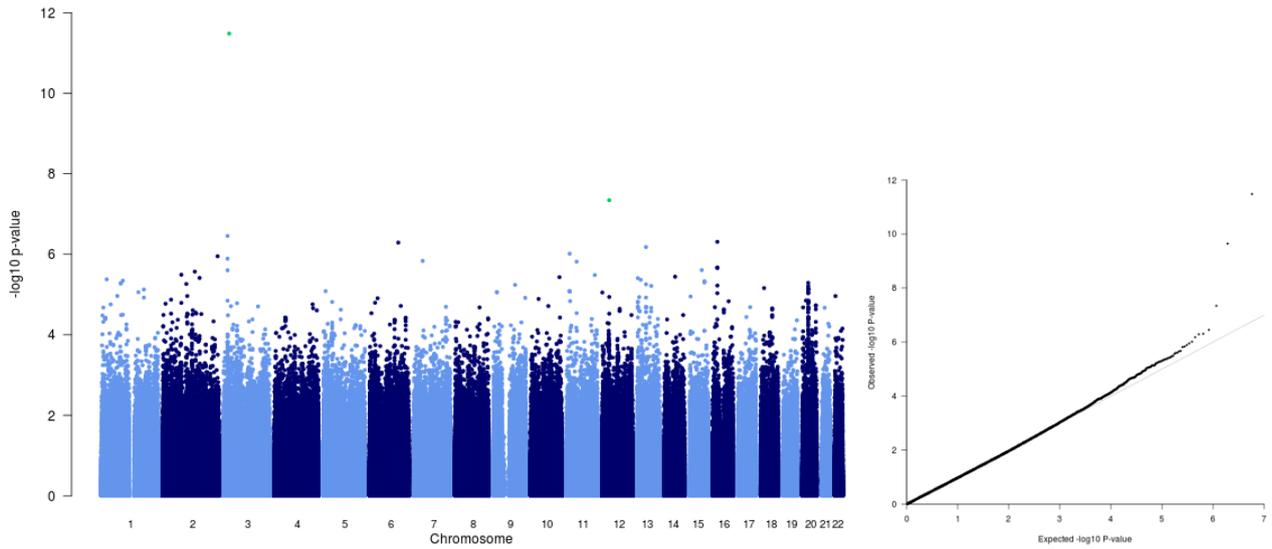
Note: LD structure is based on the 1000 Genomes European reference data.

**Figure 79** - Regional association plot showing association between EBER1 RNA levels and rs7161387, upstream of DICER1

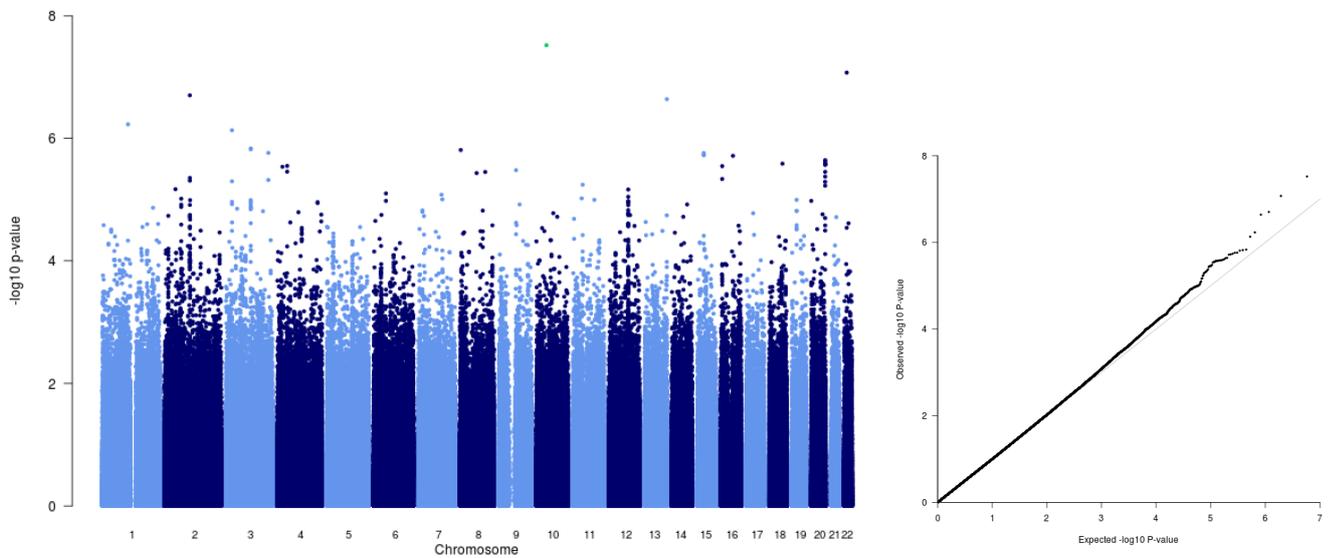


Note: LD structure is based on the 1000 Genomes European reference data.

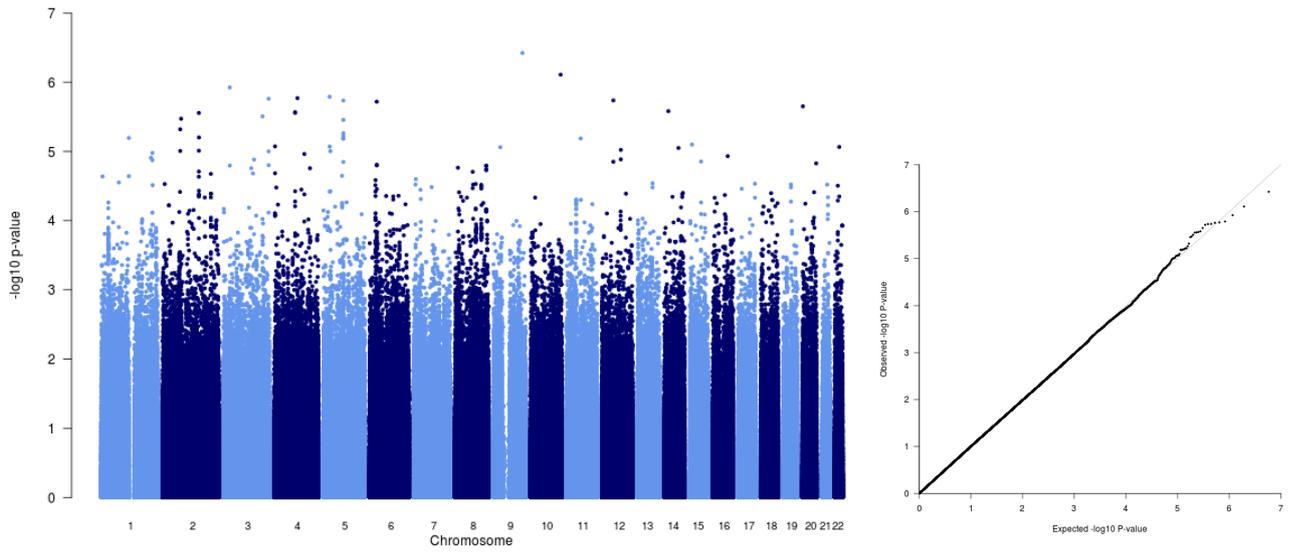
**Figure 80** - Regional association plot showing association between EBER1 RNA levels and rs12918328, upstream of RBFOX1



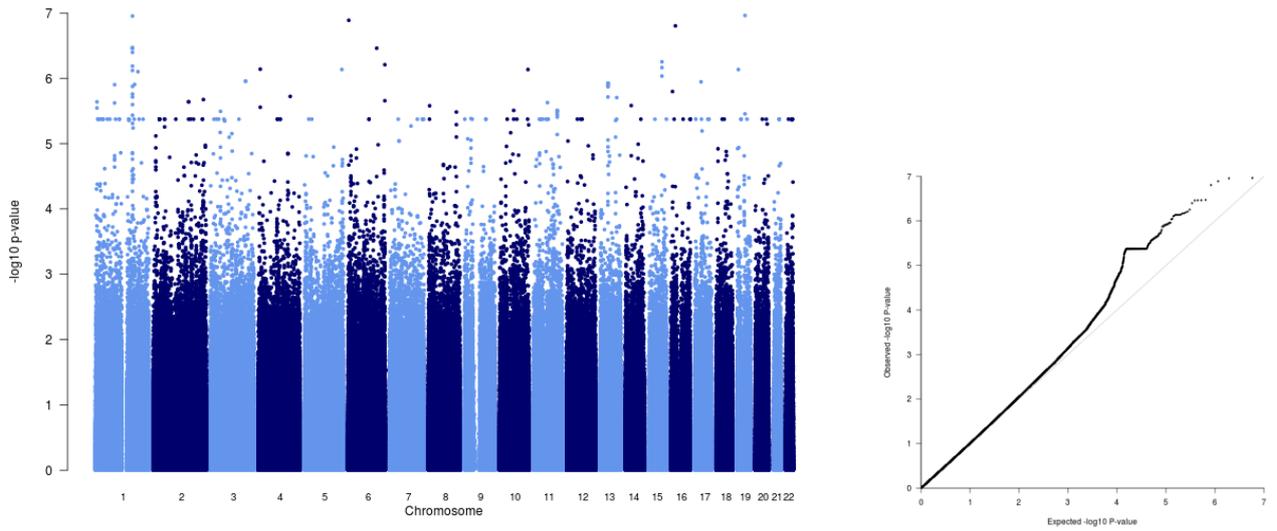
**Figure 81** - EBER 2 genome-wide association plot showing results of MRC-A/RNA-seq meta-analysis and corresponding QQ plot; negative log of P-value given by the Y axis; chromosome and expected negative of log P-value shown on the X-axis.



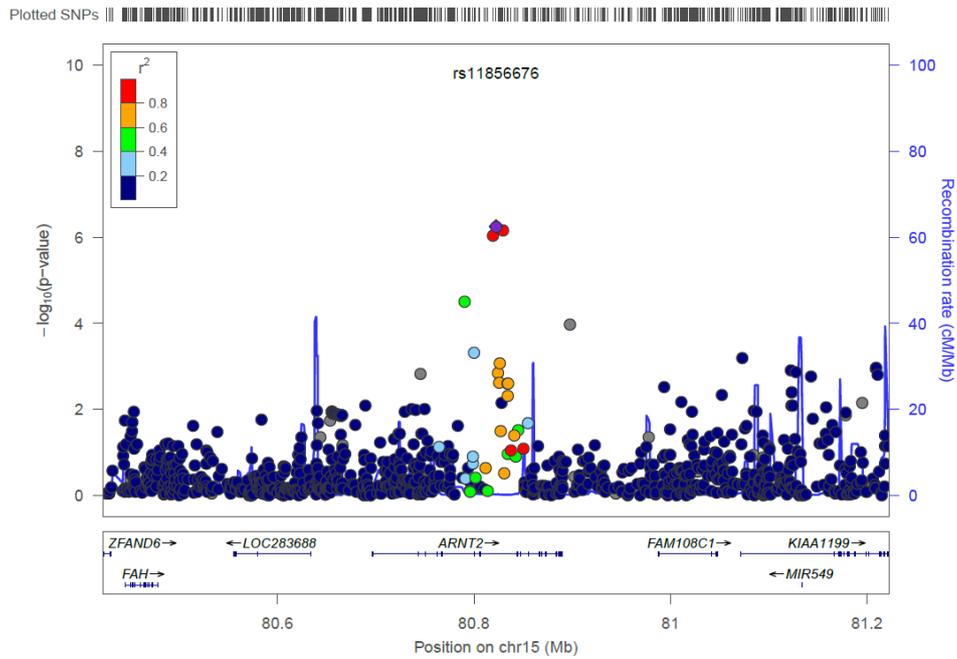
**Figure 82** – EBNA-1 genome-wide association plot showing results of MRC-A/RNA-seq meta-analysis and corresponding QQ plot; negative log of P-value given by the Y axis; chromosome and expected negative of log P-value shown on the X-axis.



**Figure 83** – EBNA-2 genome-wide association plot showing results of MRC-A/RNA-seq meta-analysis and corresponding QQ plot; negative log of P-value given by the Y axis; chromosome and expected negative of log P-value shown on the X-axis.

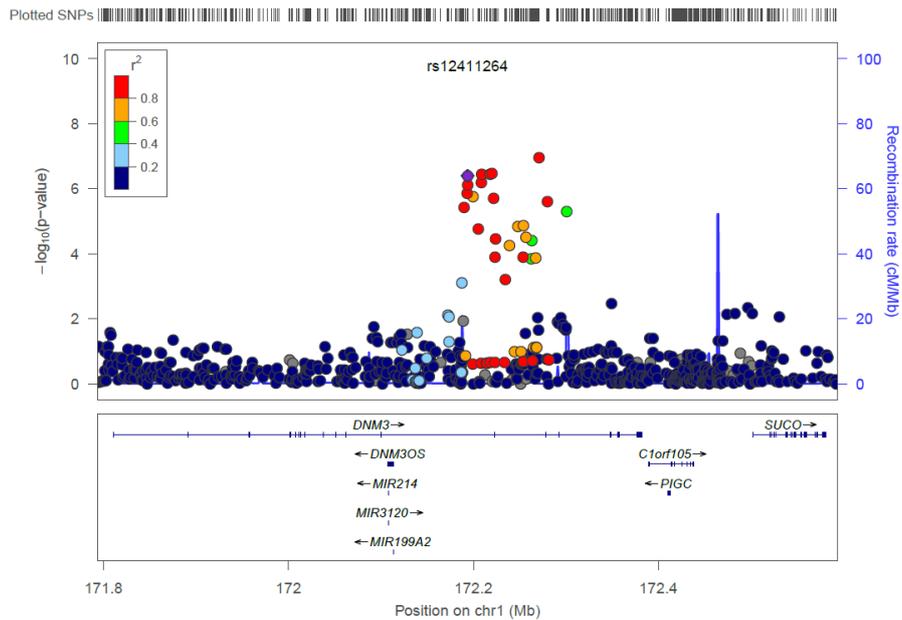


**Figure 84** – EBNA-3A genome-wide association plot showing results of MRC-A/RNA-seq meta-analysis and corresponding QQ plot; negative log of P-value given by the Y axis; chromosome and expected negative of log P-value shown on the X-axis.



Note: LD structure is based on the 1000 Genomes European reference data.

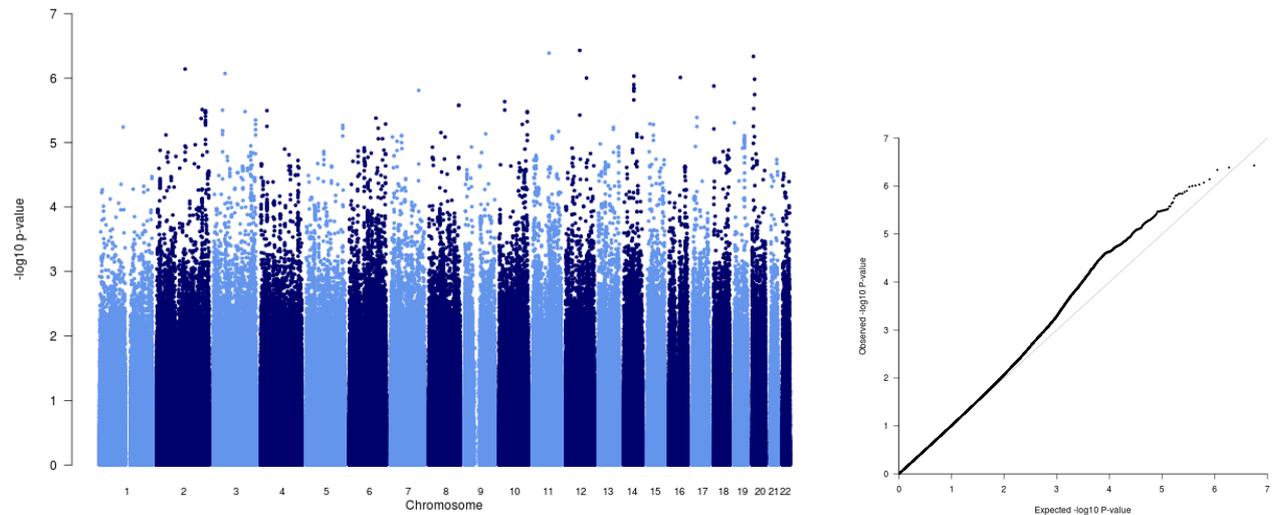
**Figure 85** - Regional association plot showing association between EBER3A expression levels and rs11856676, within ARNT2



Note: LD structure is based on the 1000 Genomes European reference data.

**Figure 86** - Regional association plot showing association between EBNA3A transcript expression and rs12411264, within DNMT3

Both the Manhattan and the QQ plot for EBNA3A indicate that inclusion of RNA-seq log-transformed expression data introduced bias, not present in MRC-A alone, and the number of associations deviates from expected.



**Figure 87** – LMP-1 genome-wide association plot showing results of MRC-A/RNA-seq meta-analysis and corresponding QQ plot; negative log of P-value given by the Y axis; chromosome and expected negative of log P-value shown on the X-axis.

In order to prioritise the candidates, regional plots were constructed for each SNP/locus listed in Table 15, and only those forming distinct, preferably continuous, association peaks were retained. After removing solitary associations and associations absent from one or more cohorts, 5 candidate loci remained. The results are presented in the Table 17. A single association reached genomewide significance, while the other remained within the suggestive range with P-values below  $1E-06$ . The direction of regulatory effects was identical for three SNPs: rs7162181, rs 11856676 and rs4238523 in all tested populations.

SNP	Chr	BP	P-value	Trait	Studies	Samples	Effect	Gene	Left Gene	Right Gene
rs6825926	4	122939162	5.01E-07	EBER1	3	408	---	NA	TMEM155	EXOSC9
rs6836544	4	122945523	8.84E-09	EBER1	3	409	---	EXOSC9	TMEM155	CCNA2
rs769243	4	122961257	9.00E-08	EBER1	3	409	---	CCNA2	EXOSC9	BBS7
rs7161387	14	94591164	5.13E-07	EBER1	3	409	---	NA	LOC730118	DICER1
rs10134473	14	94592888	8.54E-07	EBER1	3	409	---	NA	LOC730118	DICER1
rs12918328	16	6752939	1.77E-07	EBER1	3	409	++	A2BP1	LOC100131413	LOC100131080
rs7191315	16	6753913	1.18E-07	EBER1	3	409	++	A2BP1	LOC100131413	LOC100131080
rs11867159	16	6756183	4.36E-07	EBER1	1	277	+??	A2BP1	LOC100131413	LOC100131080
rs12411264	1	170460291	4.00E-07	EBNA3A	3	409	++	LOC100128178	LOC127099	LOC100131486
rs12410416	1	170460443	7.58E-07	EBNA3A	3	409	++	DNM3	LOC127099	LOC100131486
rs2289635	1	170475243	6.49E-07	EBNA3A	2	132	?+	DNM3	LOC127099	LOC100131486
rs4072117	1	170475494	3.48E-07	EBNA3A	3	409	++	LOC100128178	LOC127099	LOC100131486
rs6678749	1	170484406	3.44E-07	EBNA3A	3	409	++	LOC100128178	LOC127099	LOC100131486
rs12075079	1	170486618	3.38E-07	EBNA3A	3	409	++	LOC100128178	LOC127099	LOC100131486
rs16844255	1	170537673	1.11E-07	EBNA3A	2	342	++?	DNM3	LOC100128178	LOC100131486
rs7162181	15	78606479	9.23E-07	EBNA3A	3	408	+++	ARNT2	LOC730126	FAM108C1
rs11856676	15	78609602	5.58E-07	EBNA3A	3	408	+++	ARNT2	LOC730126	FAM108C1
rs4238523	15	78616704	6.83E-07	EBNA3A	3	409	+++	ARNT2	LOC730126	FAM108C1

**Table 17** – Candidate loci for EBV latency eQTLs from the MRC-A/RNA-seq meta-analysis

- i) “SNP” lists the rs number of the associated SNP
- ii) “Chr” and “BP” indicate the genomic position by chromosome and base pair
- i) “P-value” lists the test statistic P-values for each SNP
- ii) “Trait” indicates the associated phenotype
- iii) “Studies” informs how many cohorts had a genotype available for a given SNP
- iv) “Samples” lists the number of samples available for meta-analysis.
- v) “Effect” shows the direction of the SNP’s regulatory effect on transcript levels in each cohort
- vi) “Gene” indicates whether the variant is located within a gene and provides its name
- vii) “Left/Right Gene” list the two closest genes flanking the relevant SNP

A correlation test between EBV latency transcript and MRCA Human transcript levels revealed significant correlation between BART and human heavy chain constant gamma 1 gene (*IGHG1* - encoding immunoglobulin G1 heavy chain constant region) transcript level (FDR 0.037), and EBER1 and *SLC22A23* (encoding a transmembrane protein transporting organic ions across cell membranes) transcript level (FDR 0.013)

### 5.3.2 EBV copy number meta-analysis

Meta-analysis of EBV copy number has been conducted using four populations from the 1000 Genomes samples (n 277), and three HapMap populations from the Choy et al. (2008) study including CEPH (n 87), Yoruba (n 85) as well as additional 45 Japanese samples, which have not been previously included in any analysis in the current work or tested on their own because of the small total sample size. The genomic control factor values (Table 18) and

22 top associations from the EBV copy number meta-analysis ( $p < 2.00E-06$ ) (Table 19) are provided below.

Cohort	Lambda
1000G	1.08
Choy CEPH	1.08
Choy YRI	1.01
Choy JPT	1.04

**Table 18** - Genomic control factor for samples used in EBV copy number meta-analysis.

Moderate inflation due to population structure was present in the 1000G and CEPH samples, therefore the genomic control correction ( $-gco$  option in GWAMA) has not been applied to the copy number meta-analysis. 22 top associations for the EBV copy number meta-analysis ( $p < 2.00E-06$ ) are summarised in the table below. The second part of Table 19 looks at the heterogeneity of effects for the three SNPs genotyped in all 4 cohorts.

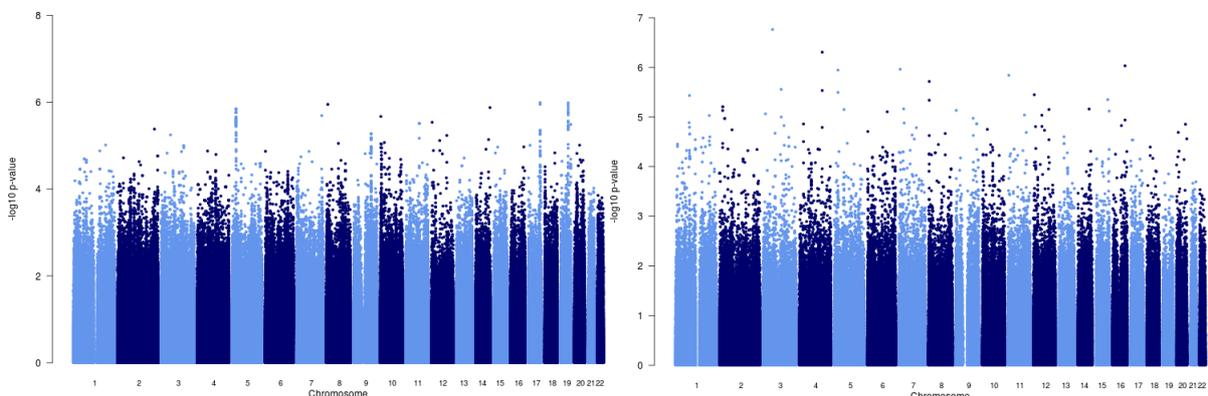
SNP	Chr	BP	p-value	Studies	Samples	Effect	Gene	Left Gene	Right Gene
rs10520897	5	23337132	1.72E-06	1	287	???		LOC100128381	LOC391771
rs13180805	5	23338554	1.72E-06	1	287	???		LOC100128381	LOC391771
rs13164226	5	23338760	1.72E-06	1	287	???		LOC100128381	LOC391771
rs13167341	5	23341873	1.72E-06	1	287	???		LOC391771	PRDM9
rs34549526	5	23351868	1.44E-06	1	287	???		LOC391771	PRDM9
rs34075772	5	23353085	1.44E-06	1	287	???		LOC391771	PRDM9
rs34803213	5	23355924	1.44E-06	1	287	???		LOC391771	PRDM9
rs35624884	5	23359151	1.68E-06	1	287	???		LOC391771	PRDM9
rs6861201	5	23360108	1.44E-06	1	287	???		LOC391771	PRDM9
rs13180619	5	23362445	1.44E-06	1	287	???		LOC391771	PRDM9
rs59556584	5	23399654	1.43E-06	1	287	???		LOC391771	PRDM9
rs1609819	8	8534826	1.12E-06	1	287	???		PRAGMIN	CLDN23
rs10873551	14	1.04E+08	1.33E-06	1	287	???		RPS2P4	KIAA0284
rs4789129	17	70583107	1.02E-06	1	287	???		KCTD2	SLC16A5
rs4513132	17	70586913	1.12E-06	4	430	-++		KCTD2	SLC16A5
rs10403833	19	45814259	1.74E-06	1	287	???	LTBP4	SHKBP1	NUMBL
rs10403467	19	45814336	1.96E-06	1	287	???	LTBP4	SHKBP1	NUMBL
<b>rs10403963</b>	<b>19</b>	<b>45814511</b>	<b>1.75E-06</b>	<b>4</b>	<b>430</b>	<b>+++</b>	LTBP4	SHKBP1	NUMBL
rs28655571	19	45816717	1.74E-06	1	287	???	LTBP4	SHKBP1	NUMBL
rs12980111	19	45818411	1.17E-06	1	287	???	LTBP4	SHKBP1	NUMBL
<b>rs8107014</b>	<b>19</b>	<b>45819305</b>	<b>1.53E-06</b>	<b>4</b>	<b>430</b>	<b>+++</b>	LTBP4	SHKBP1	NUMBL
rs11667706	19	45819366	1.33E-06	1	287	???	LTBP4	SHKBP1	NUMBL
rs11668582	19	45819653	1.03E-06	1	287	???	LTBP4	SHKBP1	NUMBL

**Table 19** – EBV copy number meta-analysis. Legend – as in Table 17. “?” in the “Effect” column indicates the SNP had not been genotype in the cohort. SNPs with identical effect across all tested cohorts shown in bold.

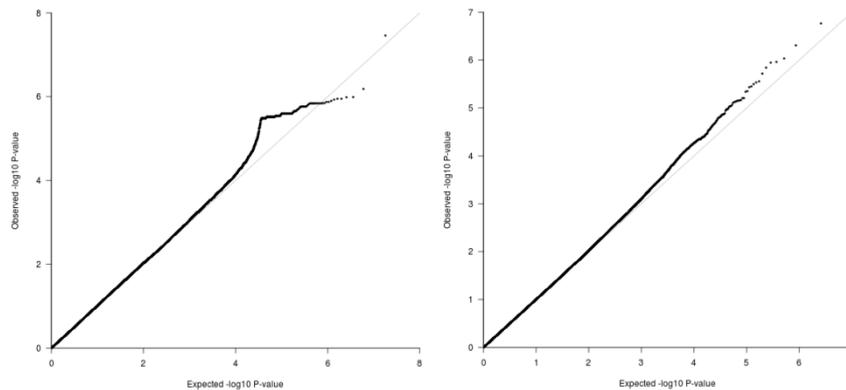
rs number	q statistic	q p-value	I <sup>2</sup>	n studies	n samples
rs4513132	4.94	0.18	0.39	4	430
rs10403963	0.70	0.87	0.00	4	430
rs8107014	0.92	0.82	0.00	4	430

**Table 19A** – EBV copy number meta-analysis – heterogeneity of effects. Legend – as in Table 14A.

After the meta-analysis, no association was statistically significant. The results mirrored those of the previous experiment conducted using 1000 Genomes samples only. Except for solitary associations, only the three previously identified candidate QTL loci (on chromosomes 5, 17 and 19) were present (Figure 88). This is not surprising as there were no significant QTLs found in the HapMap samples when analysed previously. Most variants mapped in the 1000 Genomes samples were missing from the HapMap 3 genotypes, however two SNPs on chromosome 19 (rs10403963 and rs8107014 - shown in bold) revealed identical effect across all tested samples. The third SNP that was genotyped in all the tested populations, rs4513132, revealed moderate (but insignificant) heterogeneity of effect with opposing direction of effect in two cohorts.



**Figure 88** - Legend on the next page.



**Figure 88** - Genome-wide association plot showing results from the combined MRC-A/CEPH/YRI/Japanese meta-analysis and a corresponding QQ plot (left); and results from Choy et al. samples only without MRC-A (right).

In addition, when SNPs from the EBV copy number meta-analysis were compared against the MRC-A human expression eQTLs, rs8107014 was found to be a *cis* eQTL for the transcript of a nearby gene, *SHKBP1* (P-value  $2.32E^{-06}$ , FDR 0.00072). No overlap with *trans* eQTLs have been found and the only single overlap with risk loci from the GWAS Catalog was for a non-significant association of rs12203592 (not shown in Table 19, copy number P-value  $3.68E^{-05}$ ), in *IRF4*. *SHKBP1* encodes SH3 domain-containing kinase-binding protein 1 which is a poorly characterised inflammation biomarker also implicated in apoptosis and in Acute Myelogenous Leukemia (Rao and Smith 2013, Smith and Khanna 2010, Greif et al. 2011, Azuaje et al. 2011). It has no links to EBV infection.

### 5.3.3 EBV transcript / copy number correlation

EBV copy number and expression data were available for 48 CEPH and 52 YRI HapMap LCLs grown by Choy et al and Arvey et al in two independent experiments. A Spearman's rank correlation test was implemented to test copy number versus all 83 EBV transcripts quantified from the RNA-seq studies' data (Figure 89). When significant correlations



## 5.4 Discussion

### 5.4.1 Latency eQTLs

Five eQTLs were resolved using meta-analysis. Three involved EBER1. The most significant (rs6836544) spanned a region containing *CCNA2*, a housekeeping gene and a general regulator of the cell cycle, and *EXOSC9*, a component of the exosome complex which processes and degrades RNA (Akiyama 2014). Another, this time a suggestive, association was found for a locus ~30kb downstream of the gene encoding *DICER1*, a protein with endoribonuclease/helicase properties responsible for synthesis of small interfering RNA and post-transcriptional silencing via RNA-induced silencing complex (RISC) (Watson et al. 2008). This complex is pivotal against viral infection by degrading viral RNA and many viruses have evolved ways to evade RISC (Pumplin and Voinnet 2013). Finally, the EBER1 eQTL previously identified in the MRC-A samples and located in *RBFOX1* (also known as *A2BPI*), was replicated in the meta-analysis, albeit not at the level of genomewide significance. The two remaining suggestive eQTLs were associated with EBNA3A transcript levels. One was located on chromosome one within the ORF of *DNM3*, encoding Dynamin 3. Dynamin 3 is a GTPase which mediates vesicle formation from cellular membranes. *DNM3* is maximally expressed within human CNS, and it has been suggested as potentially relevant to neurodegenerative diseases (Romeu and Arola 2014). On the other hand, *DNM3* has been reported to act as a tumour suppressor in hepatocellular carcinoma (Inokawa et al. 2013). The gene was also shown to be over-expressed in a rare type of T-cell lymphoma, characterised by localization of malignant cells to the skin, called Sezary syndrome (Wang et al 2011). In addition, a recently conducted ChIP-Seq experiment indicated *DNM3* can interact with EBNA1 (Sato and Tabunoki 2013).

The last of the candidate eQTLs was present on chromosome 15. This locus contains *ARNT2* and encodes aryl hydrocarbon nuclear translocator 2, a basic-helix-loop-helix transcription factor expressed predominantly in neurons (Maltepe et al. 2000, Drutel et al. 1999) *ARNT2* forms dimers with other factors like HIF $\alpha$  and modulates the cell's response to environmental stimuli (Hankinson 1994, McIntosh et al 2010). The meta-analysis overall decreased the significance of the MRCA findings. Additionally the viral/human eQTL overlap investigation did not yield any new insights. A potentially interesting correlation was found between BART expression levels and immunoglobulin gene *IGHG1* expression levels within the MRCA samples. *IGHG1* product has been implicated in immune response to bacterial and viral infections and influencing course of allergies, autoimmunity and malignancy development (Oxelius and Pandey 2013).

#### **5.4.2 Copy number**

Because of the small sample size of the HapMap populations provided by Choy et al. (2008), the copy number meta-analysis largely mirrored the results obtained using the 1000 Genomes panels, albeit with a slight overall reduction of significance levels. An interesting finding was the same direction of effect across all populations for two SNPs within the latent transforming growth factor beta binding protein 4 gene, which adds credibility to the association.

#### **5.4.3 EBV transcript and copy number correlation**

The significant association between viral copy number and the majority of tested latency and lytic transcripts with clear-cut opposite effects was a notable finding. This effect has some

important implications for the current project. Firstly, it indicates that the correlation between EBV expression and copy number is not random and increased copy number is characteristic of increased lytic expression. This is most likely a result of the more efficient rolling circle replication, producing more virions and leading to a higher re-infection rate. Secondly the effect spans different populations and was still evident in a mixed 50% Yourban 50% Caucasian cohort. Finally, it has to be noted that the investigated LCLs had been grown in two independent experiments. Still, the variation in EBV copy number observed across a panel of cell lines in one experiment almost perfectly explained the variation in lytic versus latent expression in the same set of cell lines from another assay. This suggests that the proportion of lymphoblastoid cells expressing lytic transcripts and replicating lytically in each cell line is at least partly under genetic control. The finding corroborates the conclusions of other authors. Arvey et al. (2013) reported that LCLs are characterised various levels of lytic expression which correlated with EBV copy number and Caliskan et al. (2011) found evidence for genetic determinants of EBV copy number in LCLs. Although the causal effect has yet to be determined, it is plausible that both factors act synergistically. Elevated lytic replication increases the viral load at the same time promoting re-infection, which in turn can also elicit lytic response in latently-infected B- cells. Pinpointing specific genetic determinants of this phenomenon and investigating them functionally should explain the causal factor behind the observed correlation. As for the copy number association/GWAS Catalog overlap, the association of rs12203592 located within *IRF4*, a gene influencing skin and eye colour (GWAS Catalog) was not significant. However it is worth noting that by upregulating the NF- $\kappa$ B pathway through EBNA2, EBV is known to increase *IRF4* expression, and *IRF4* knock-down experiments in LCLs decreased cell division (Shaffer et al. 2009).

## 5.5 Conclusion

All of the significant and suggestive associations found in the meta-analysis of MRC-A and RNA-seq samples were either associations involving one to three SNPs, or located within genes with no immediate link to EBV biology. Some of those genes, such as *ARNT2*, *DNM3* or *RBFOX1* are characterised by a restricted expression pattern across human tissues, with little relevance to lymphoid disorders. Meta-analysis of the copy number did not alter the main results of the 1000 Genomes samples copy number assay, however it brought an interesting finding indicating that two suggestively associated SNPs within *LTBP4* have an identical effect across all 4 tested cohorts. These were also the only two SNPs within the association peak, whose genotypes were available for all samples and *LTBP4* is the only one of the three suggestive copy number associations which bears a direct relevance to viral infection. This favours the possibility that endogenous TGF- $\beta$  secreted by activated B-cells as an autocrine growth inhibitor (Lagneaux et al. 1997, Kehrl et al. 1986), may indeed be influencing EBV infection *in vitro* by promoting lytic replication and therefore increasing chances of successful re-infection. If real, this effect would likely be active at the very initial stage of B-cell transformation and LCL outgrowth since mature LCLs are largely insensitive to TGF- $\beta$ . The fact that LCLs characterised by higher EBV load also have a uniformly greater propensity for lytic expression adds support to this hypothesis.

## **Chapter 6. MRC-A latency eQTL follow up**

### **6.1 Introduction**

In order to validate two of the most significant EBV latency eQTLs identified in the MRC-A cohort (rs17158630 for EBNA1,  $1.10E^{-08}$ ; and rs17339199 for EBNA3A,  $2.12E^{-08}$ ; both described in Chapter 3) further work was conducted with the aim of replicating their association with both the associated viral transcripts as well as human transcripts CD83 and IL12B (for which SCAN identified a regulatory effect of the candidate SNPs) respectively. The experimental design was based on recruitment by genotype for informative SNPs and investigation of association with gene expression in three populations of B-cells: A) primary unstimulated B cells B) primary B cells activated by a cytokine cocktail mimicking antigen-driven activation C) LCLs established from the same individuals and sampled at least two different time points – during the outgrowth and once the cultures become stable. The latter would allow to assess if the effects of candidate eQTLs change over time or are active only during the initial stage of transformation.

### **6.2 Aims of the chapter**

1. To validate the association for two candidate latency eQTLs (a) rs17158630 associated with EBNA1 expression, and a putative eQTL for RASFGF1A and CD83; (b) rs17339199 associated with EBNA3A, and possible eQTL for FKBP6 and IL-12B.
2. To compare associations between the two latency eQTL SNPs and human transcripts in naïve B-cells, B-cells stimulated with a cytokine cocktail mimicking antigen-driven activation and LCLs in order to check whether the association holds under specific conditions.

3. To investigate the associations of the two candidate eQTLs with EBV transcript at different time points following EBV immortalisation and establishment of LCLs to investigate the effect of EBV transformation on the eQTLs.

## 6.3 Results

### 6.3.1. Study design

A total of 39 healthy volunteers of European ancestry were recruited by specific eQTL candidate's genotype from the Oxford Biobank to be maximally informative for the two SNPs of interest (Table 20).

<b>Candidate eQTL:</b>	<b>rs17158630</b>	<b>rs17339199</b>
eQTL for	EBNA1	EBNA3A
Located in	RASGEF1A	FKBP6
SCAN eQTL for	CD83	IL-12B
<b>Blood Samples:</b>	<b>39</b>	<b>39</b>
wild-types	24	23
heterozygotes	12	12
homozygotes	3	4
<b>LCL Samples:</b>	<b>23</b>	<b>23</b>
wild-types	15	15
heterozygotes	6	5
homozygotes	2	3

**Table 20** – Summary of the samples for the MRC-A eQTL study follow-up;

- i) rs numbers of the two candidate eQTLs chosen for replication are listed in the top row;
- ii) the corresponding associated EBV latency transcript is listed in the second row;
- iii) third row lists genes in which the eQTL candidates are located ;
- iv) human genes whose expression was also associated with the candidate EBV eQTL (with  $p < 9E-05$ ) are listed in the fourth row; human eQTL data has been sourced from the SCAN database;
- v) rows 5-8 give the total number of individuals whose peripheral B-cells were used in the follow-up, and number of samples by genotype;
- vi) rows 9-12 give provides information on the subset of individuals whose blood was used to generate LCLs for the purpose of the follow-up, and how many samples were available for each genotype;

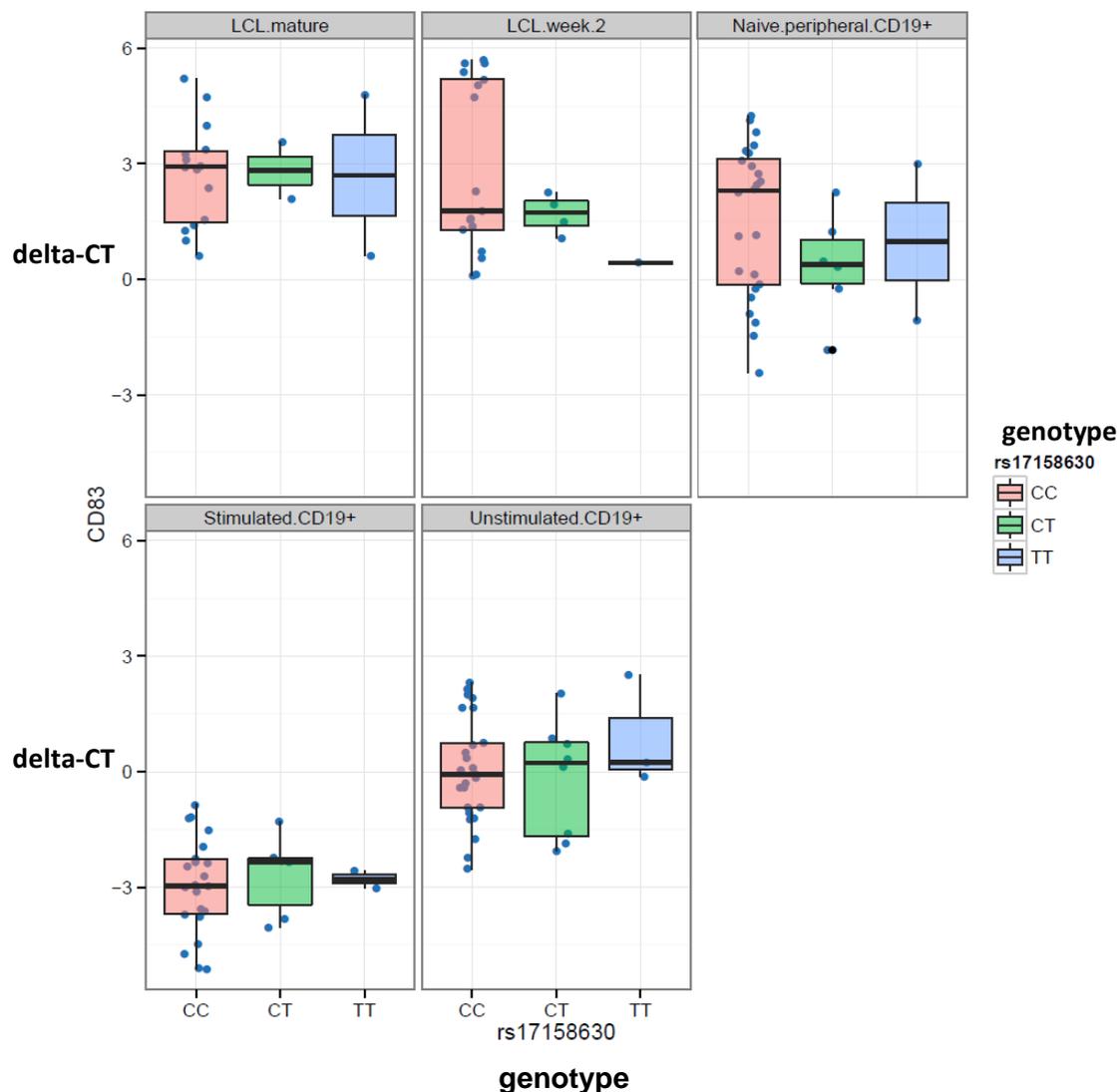
Following venesection, PBMCs were isolated using a ficoll gradient and primary CD19+ B cells purified and either incubated with a growth-stimulating and activation-eliciting cytokine cocktail (Wagner et al. 2004, He et al. 2004) composed of CpG ODN, IL-4 and CD40 ligand (n=39) or incubated for 24 hours with no stimulants, and then harvested (n=39). In addition, whole blood was sent to [Public Health England Cell Culture Service in Porton Down, Wiltshire UK] in order to establish a LCL for each volunteer (n=23 at the time of completing this thesis). Lymphoblasts were harvested twice, during the 2<sup>nd</sup> week of cell line outgrowth (denoted “LCL week 2”) and from fully grown and stable cell lines (5<sup>th</sup> week of outgrowth) (denoted “mature LCL”).

All the RNA samples were converted into cDNA and assayed for transcript expression levels using TaqMan qPCR probes. All experiments were conducted in duplicate. EBV latency transcripts were assayed in LCL samples only. EBV infects only a small proportion of peripheral memory B-cells (0.5 to 50 per million) (Thorley-Lawson 2001) and remains latent shutting down the expression of its genes (except for EBERs and EBNA1 – Latency I), hence EBV transcript quantification is not possible in RNA derived from peripheral blood B-cells. The results of the follow up are presented below. The expression levels of *IL-12B* and *RASGEF1A* were comparatively low compared to *CD83* and the housekeeping gene, *OAZ1*. This effect was particularly evident in cDNA derived from peripheral blood B-cells. In several samples qPCR failed to amplify *IL-12B* and *RASGEF1A*. *FKBP6* expression levels were very low and could be only obtained for 2 naïve and 2 unstimulated samples as well as 8 LCL samples. Consequently, *FKBP6* had to be excluded from further analysis. Expression values were quantified as Ct and normalised to the housekeeper *OAZ1* expression levels (delta-Ct) before analysis. Because only 2 to 3 homozygotes were available for the tested loci, it was necessary to count them together with the heterozygotes when conducting statistical tests. Consequently, for each investigated transcript, two sample t-tests were

conducted that compared the expression in terms of meaned delta-Ct of all samples possessing the minor allele vs those with no copy of the minor allele (wild-type genotype). A summary of samples available for the test is given in the appendix by Table A-2.

### 6.3.2. rs17158630, candidate eQTL for EBNA1

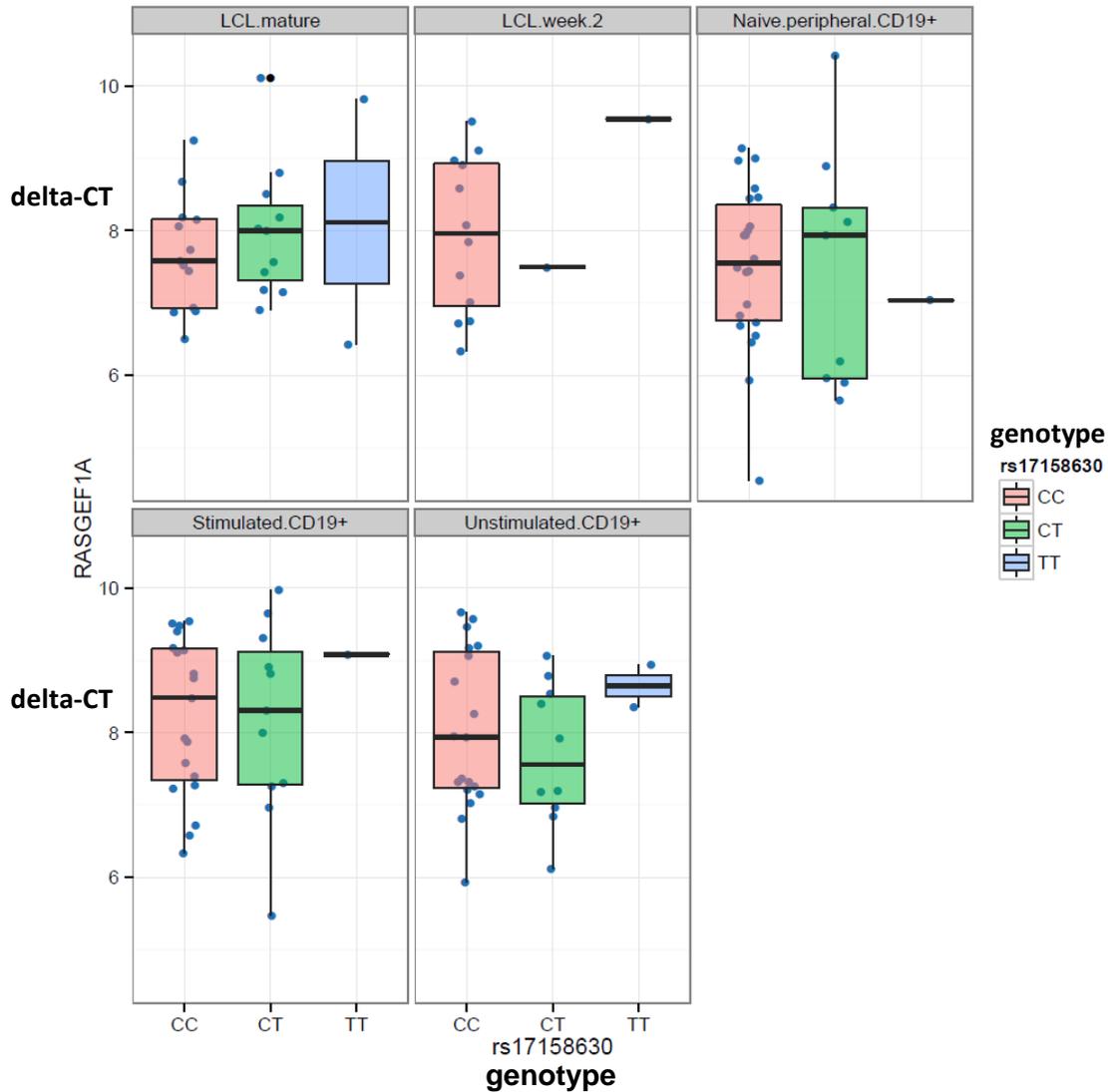
*CD83* expression by genotype was analysed for primary B cells as well as in LCLs (Fig 90).



**Figure 90** - Box-plots show *CD83* expression (normalised to *OAZ1* housekeeper and measured in delta-CT) by rs17158630 genotype in all experimental sets. Sample genotypes are coloured red (wild-type), green and blue (for minor allele homozygote), as indicated by the legend on the right; box-plot whiskers indicate the min and max d-Ct value; the actual box-plot borders show the first and the third quartile; the vertical bar indicates the median.

No association with genotype was observed, either in the primary B cells or following EBV immortalisation at the early and late time point. However a two-sample t test indicated a statistically significant (P-value < 0.001) difference between the naïve and stimulated B-cell samples as well as unstimulated vs stimulated with higher expression levels after stimulation. The d-CT means were 1.26 (Naive) 0.41 (Unstimulated) and -2.59 (Stimulated), and the SD approximately 1.85 for all three. No significant difference in CD83 expression was present between the two LCL time points.

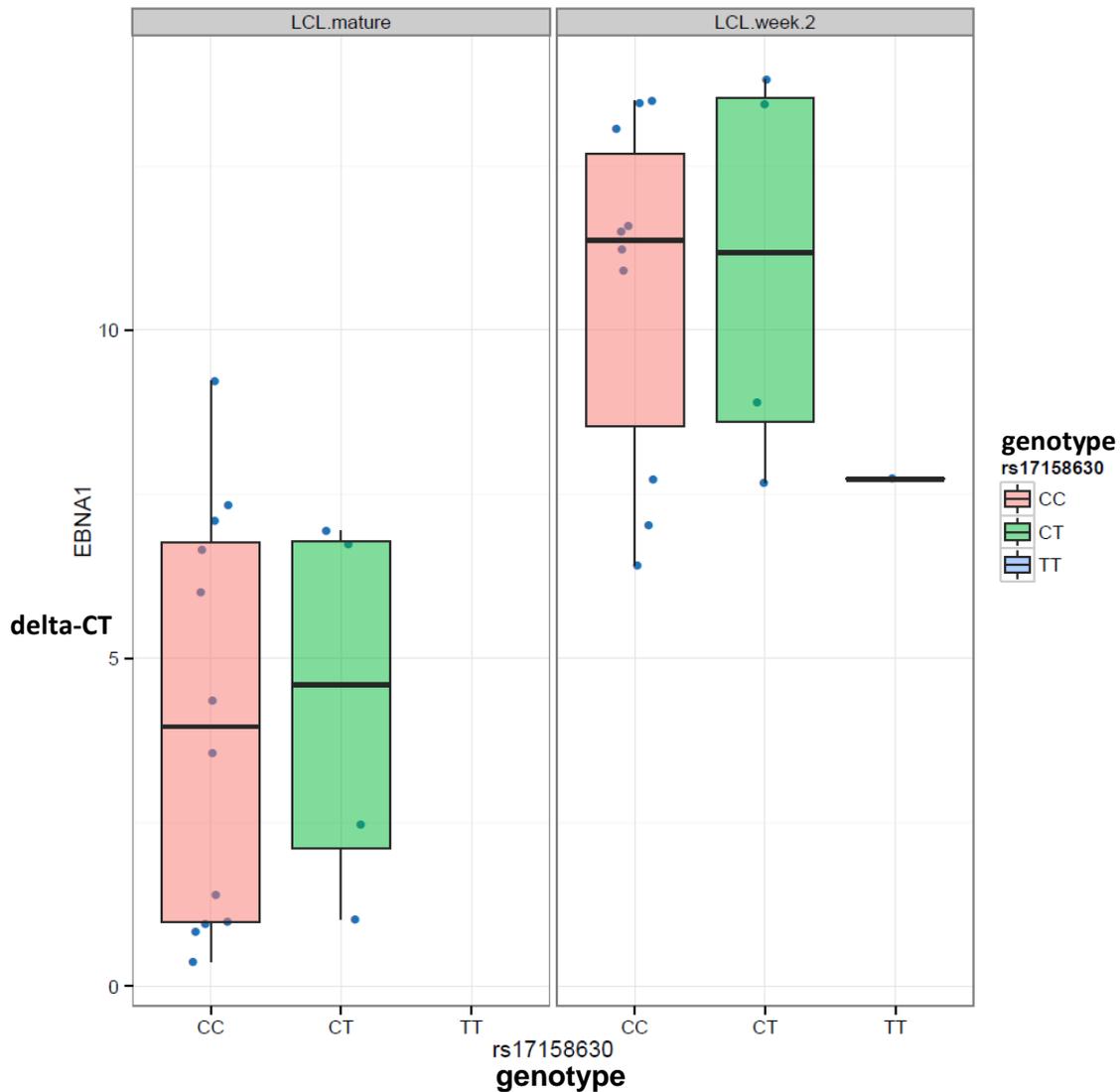
***RASGEF1A*** - *RASGEF1A* expression could not be obtained for all peripheral B-cell samples, however the transcript amplified successfully in all LCL samples. Analysis of *RASGEF1A* expression by genotype (Figure 91) indicated no statistically significant difference between the tested samples. Statistically significant difference in expression levels was however found between the peripheral naïve B-cells and stimulated B-cell samples (lower expression present after 24h stimulation, P-value < 0.01). The means were respectively 7.38 and 8.33, and the SDs 0.97 and 1.14.



**Figure 91** – Box-plots of RASGEF1A expression levels (delta-Ct values) by rs17158630 genotype in all experimental sets. Sample genotypes are coloured red (wild-type), green and blue (for minor allele homozygote), as indicated by the legend on the right; box-plot whiskers indicate the min and max d-Ct value; the actual box-plot borders show the first and the third quartile; the vertical bar indicates the median.

**EBNA1** - The viral EBNA1 latency transcript was quantified in the LCL samples only (Figure 92). Comparison of expression level means by genotype (wild type vs others) revealed no statistically significant difference between the two groups in both sample sets. The test indicated that there was a significant difference in mean expression levels of EBNA1 when the early culture replicates were compared to fully grown LCLs. EBNA1 expression levels were higher (P-value < 0.001) in the viable and established LCLs (lower delta-Ct).

Means were 10.5 (week 2) and 4.3 (mature LCL), while the SDs were high and respectively, 2.69 and 3.03.

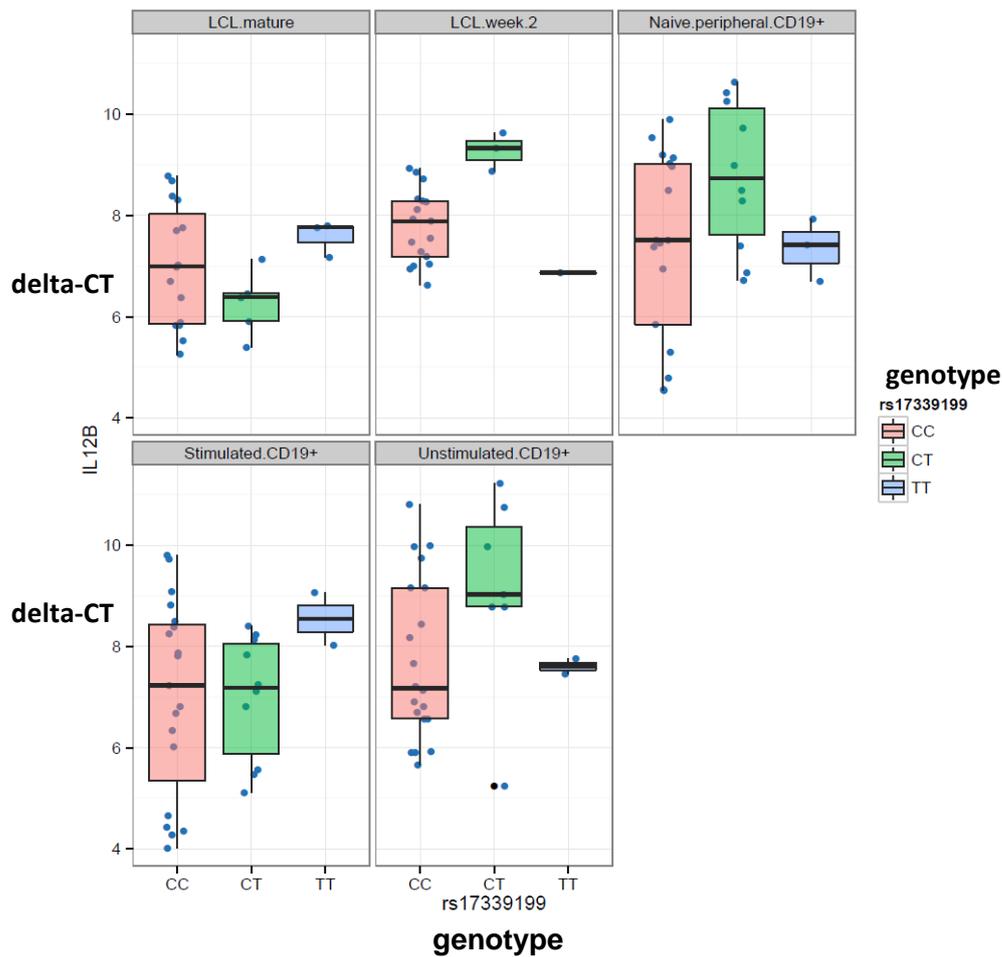


**Figure 92** – EBNA1 expression levels(delta-CT) by rs17158630 genotype. Comparison of expression levels at the end of the second week of LCL outgrowth vs the end of outgrowth (mature LCL). Sample genotypes are coloured red (wild-type), green and blue (for minor allele homozygote), as indicated by the legend on the right; box-plot whiskers indicate the min and max d-Ct value; the actual box-plot borders show the first and the third quartile; the vertical bar indicates the median.

### 6.3.3. rs17339199, candidate eQTL for EBNA3A

*IL-12B* - The expression-by-genotype two sample t-tests revealed no statistically significant differences in mean transcript abundance between the rs17339199 wild-types and other

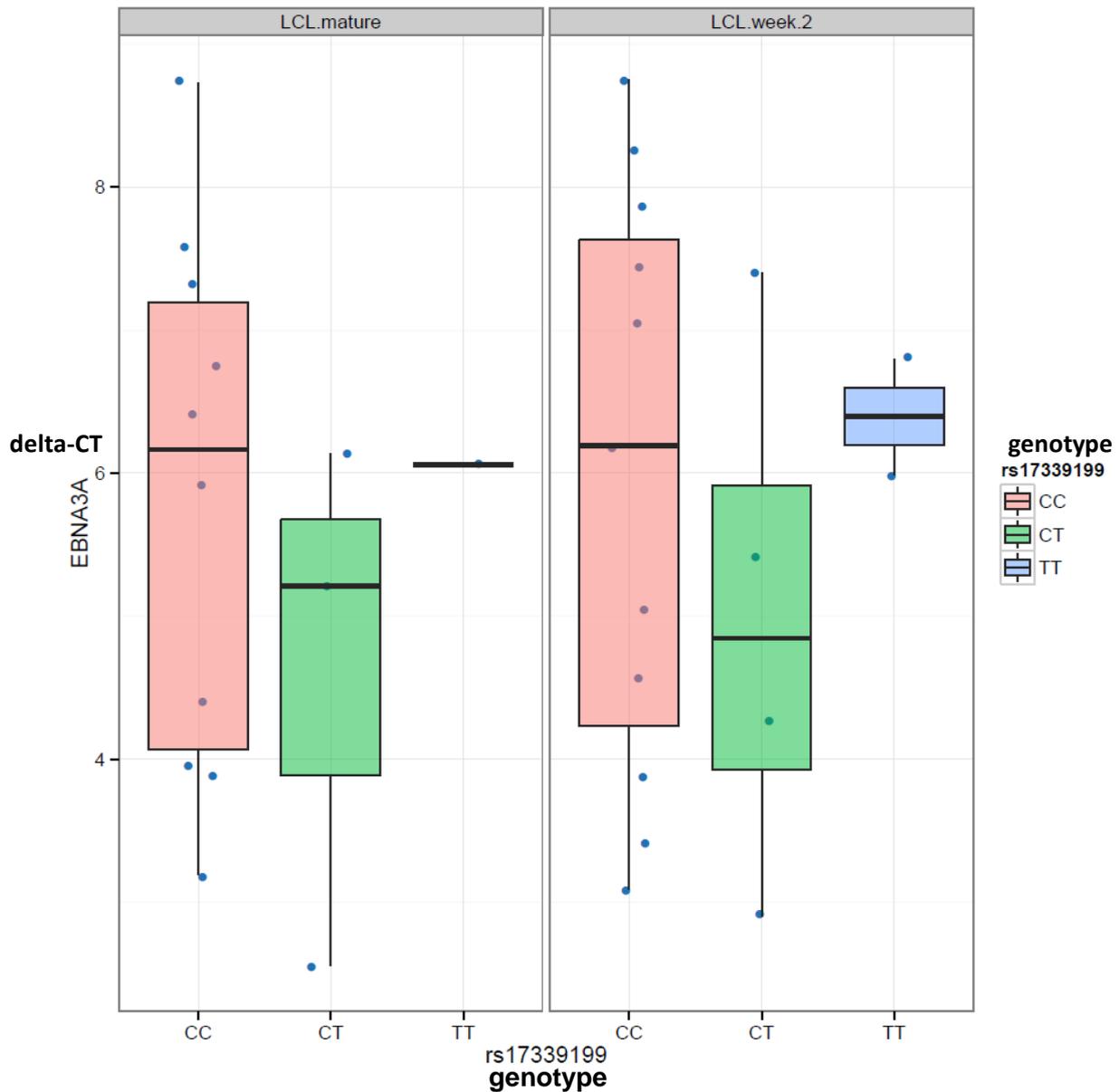
samples within each experimental sample set (Figure 93). Across the different sample sets, the mean expression levels differed significantly when unstimulated B-cells were compared to stimulated samples, with higher expression levels after stimulation, (P-value < 0.05). Respective means were 7.9 and 7, SDs 2 and 1.8. The two time-point replicates of the LCL samples also differed significantly (higher expression levels in fully grown out LCL, P-value < 0.01). Means were 7.9 (week 2) and 6.9 (mature LCL), the SDs 0.8 and 1.



**Figure 93** - Box-plots of IL-12B expression levels (delta-Ct values) by rs17339199 genotype in all experimental sample sets. Sample genotypes are coloured red (wild-type), green and blue (for minor allele homozygote), as indicated by the legend on the right; box-plot whiskers indicate the min and max d-Ct value; the actual box-plot borders show the first and the third quartile; the vertical bar indicates the median.

**EBNA3A** - There were no statistically significant differences in EBNA3A expression levels between mature and week-2 LCLs. Also, when wild-type individuals were compared to the

others, the rs17339199 genotype did not have any effect on the transcript abundance in either sample set.



**Figure 94** - EBNA3A transcript levels by rs17339199 genotype. Comparison of expression levels at the end of the second week of LCL outgrowth vs the end of outgrowth (mature LCL). Sample genotypes are coloured red (wild-type), green and blue (for minor allele homozygote), as indicated by the legend on the right; box-plot whiskers indicate the min and max d-Ct value; the actual box-plot borders show the first and the third quartile; the vertical bar indicates the median.

## **6.4 Discussion**

No significant association was seen between the genotypes of interest and any of the investigated human and EBV latency transcripts. This undermines the significance of the two candidate EBV eQTLs identified in the MRCA cohort and does not support the human eQTL findings from the SCAN database. The qPCR assays yielded expressions with high SD values, which makes comparison difficult, particularly in small sample sets. All tested transcripts had also low or non-detectable expression levels, except for CD83. This further reduced the power of the follow up to detect associations as more samples had to be removed before testing.

Slightly higher expression levels of immunity-related genes, CD83 and IL-12B were observed in stimulated B-cells when compared to unstimulated samples. This could be attributed to the activating effect of the cytokine IL-4, CD40L and CpG ODNs described in the literature (Wagner et al. 2004, He et al. 2004, Hasbold et al. 1999). It is expected that some genes should become more stably expressed once the LCL culture becomes fully grown and viable. This effect, however, was not consistent and could be statistically confirmed only for IL-12B, RASGEF1A and EBNA1. The difference was most notable for the EBNA1. This is likely because EBNA1, which associates itself with the viral episomes and participates in viral replication, should reflect the increasing viral copy number in the fully established growing LCL cultures.

## **6.5 Conclusion**

The findings of the follow up suggest that the results of the eQTL latency assay in the MRC-A panel should be interpreted with great caution. No statistical evidence of association could be found for the two candidate eQTLs and the two tested EBV latency transcripts. The sample size was however very small, in particular the number of homozygotic genotypes

used in the follow-up experiment was between 0 and 2, and only 4 to 5 non-wild type genotypes were available to re-investigate. The associations also indicated that the results from the public eQTL databases like SCAN should be interpreted with caution.

## **7. General Discussion**

### **7.1 Overview of the results**

The aim of this project was to conduct an extensive EBV eQTL search encompassing all transcript precursors of latency proteins, which are pivotal in the process of viral transformation of B-cells and the reason why EBV contributes to lymphoid and epithelial tissue tumourigenesis, as well as copy number QTLs. Despite its ubiquitous prevalence and links to multiple human disorders, the virus and its infection have never become a subject of GWAS, although some authors suggested that human genetic polymorphisms could influence EBV expression and copy number (Rubicz et al. 2013, Caliskan et al. 2011, Mohr 2010). The evidence they presented was based on selected viral phenotypes investigated in a small sample size and without additional replication in independent experiments. Consequently, this project aimed to fill that gap and investigated a combination of EBV phenotypes in the context of human genomic diversity with the aim to address the broad issue of global regulation of EBV expression and load by human SNPs. The assays were conducted across multiple cohorts of LCLs in order to replicate and validate the findings. LCLs are a reasonable model for this purpose as their phenotype is essentially that of an EBV-infected naïve B-cell.

An important limitation of the current study was the small sample size of most investigated LCL cohorts combined with their ethnic heterogeneity. Also the largest panel utilised for the main latency eQTL assay consisted of approximately 300 related (sibs) individuals, which decreased the power to detect associations. Altogether, these factors limited the power to identify significant associations and made replication of results challenging.

Both EBV latency eQTL and copy number QTL assays provided several candidate loci. Within the largest MRCA panel three out of five significant latency eQTLs had also been reported as human eQTLs for cytokines and receptors important for immune response, namely *CD83*, *IL12B* and *TNFRSF1B*, all of which had been linked to EBV-related autoimmune or lymphoproliferative diseases. Although most of these associations were subsequently not replicated in the HapMap YRI and CEPH panels from Arvey et al. (2013) study, other immunity-related cytokine and cytokine-receptor genes appeared suggestively (*IL2RA*) or significantly associated (*TNFSF13B* and *CSMD1*). These included two loci (*TNFSF13B* and *CSMD1*) harbouring significant, although different SNPs, in both CEPH and YRI samples. The abundance of immune response genes among the eQTL findings was made more interesting by the fact that the most significantly associated SNP within *IL2RA* had been previously identified as a risk locus for type 2 diabetes, but more importantly, risk loci for MS and RA were located just 25kb upstream within *IL2RA* itself.

In summary, several latency eQTL candidates, biologically-relevant to EBV biology and disease, have been identified by the project. Lack of replication made it difficult to prioritise the findings in terms of credibility. Their significance was also undermined by the fact that all of the statistically significant associations involved solitary or very few SNPs within any given locus. In addition, eQTL candidates from the largest MRCA cohort were located in loci with no direct biological links to EBV or lymphatic tissue, and the regulatory effects of the two top MRCA latency eQTLs (rs17158630 for EBNA1, and rs17339199 for EBNA3A) have not been replicated in the replication experiment conducted on Oxford Biobank samples. Likewise, the association of rs17158630 and rs17339199 to human transcripts, reported by the SCAN eQTL database, has also failed to replicate, however very small sample size has to be taken into account. The meta-analysis of MRCA and RNA-seq eQTLs did not increase the significance of the candidates nor provided new findings. Also, no informative associations

were provided by looking at the overlap of candidate latency eQTLs with the MRCA panel human eQTLs as well as risk loci from the GWAS Catalog.

It is however still possible, that some of the findings are genuine. Recently, several authors have reported genetic variants in immune response cytokines and TFs directly influencing the risk of EBV-positive disorders of the lymphoid tissue or epithelial cells (Salzer et al. 2012, Linka et al. 2012, Huang et al. 2011, Hildesheim and Wang 2012). Efficient expression of these pro- and anti-inflammatory cytokines by immune cells likely affects the course of such conditions, particularly PTLD (Schneider 2013). It is evident that genetic alterations to genes directly involved in T-cell and B-cell maturation and immune response, particularly those encoding interleukins, interferons or TGF- $\beta$ , contribute to the outgrowth of EBV-positive lymphoproliferative disorders and lymphomas, and the universal hallmark of those susceptibilities is high viral load characteristic of EBV viraemia during lymphoproliferation (Schneider 2013, Akay et al. 2014).

In this context, one of the most important findings of the project has been the confirmation of the close link between EBV expression and EBV copy number, an observation supported by previous studies (Ruf et al. 2014, Arvey et al. 2013). Two conclusions can be drawn from the results. Firstly, high EBV copy number in LCLs is most likely the product of lytic replication variable in a subset of cells as increased expression of most latent transcripts is inversely correlated with the viral load. It is unlikely that in the same set of LCLs grown in two independent experiments, a significant positive and negative correlation to viral load would be uniformly seen only for the lytic and only for the latent transcripts, respectively. Copy number could alternatively increase lytic expression. The study by Arvey et al. 2013 supports also supports this relationship without providing a definite answer as to mechanisms of causality although markedly high levels of IL-10 have been noted for the lytically replicating LCLs. The authors identified different LCLs can be characterised by varying ratios of lytic

and latent expression and higher EBV copy number significantly correlates with the former. The second important implication is the fact that the degree of lytic replication and consequently also EBV load is likely affected by the genetic variation within LCLs, which is indicated by the significant correlation for most transcripts and copy number in the same LCLs, but grown independently. If the tested cells were grown in a single experiment, it could have been that an environmental factor affected both EBV copy and lytic expression at the same time. Spearman's rank test indicated however that in two independent experiments the same cell lines were characterised by a negative EBV latency to copy number correlation. This regularity observed in independent subclones of the LCLs grown in different labs may indicate genetic control (Houldcroft et al. 2014). This may also explain why LCLs exhibit a similar copy number across multiple replicate cultures and much higher inter-individual variation (Caliskan et al. 2011) and also why certain LCLs are repeatedly more difficult to establish and grow than others.

Results of the copy number assay and transcript correlation put the copy number QTL assay results in a more meaningful context. Out of the three suggestive association peaks identified by the largest assay in the project, the one located in *LTBP4*, a TGF- $\beta$  transporter protein, merits particular attention because of the important role of TGF- $\beta$  in lytic reactivation of EBV in infected B-cells. BL cell lines treated with phorbol esters secrete active TGF $\beta$  and enter lytic cycle and anti-TGF $\beta$  antibodies increase lytic early antigen expression by 75-85%, which suggests that viral productive cycle is partly regulated by active endogenous TGF $\beta$  production (di Renzo et al. 2006). This effect has been absent from LCLs in Latency III, possibly because of expression of LMP1 or EBNA2 blocks the responsiveness to TGF- $\beta$  signalling (Inman and Allday 2000, Campion et al. 2014). It is possible however that it is active during other stages of EBV infection both *in vivo*, for instance at the Latency II germinal centre stage, and also *in vitro*, during B-cell transformation and early LCL

outgrowth when the virus is only beginning to affect cellular biology. Potential role of LTBP4 in attenuation of the effect of endogenous TGF- $\beta$  signalling on apoptosis and EBV reactivation could be the reason for association and it is worth of further investigation.

## **7.2 Future Research**

There are several ways in which the EBV QTL findings could be further investigated. The most straightforward would involve the use of the statistical tools which have already been utilised for the purpose of current project. Firstly, all tests included in the meta-analysis could be re-done with appropriate measures taking into account the population structure, for instance GWAMA's Mantra feature. In this way the association results could be refined and no P-value inflation would be present despite testing cohorts from different ethnic backgrounds (Morris 2011).

The optimal approach would be to first apply a robust conservative normalisation method, preferably quantile normalisation, to the expression and copy number values in order reduce the number of false positives and enforce normal distribution. Then, to employ a mixed linear model approach for all the conducted association tests which would effectively control for population structure, familiarity as well as cryptic relatedness at the same time (Zhang et al. 2010, Eu-ahsunthornwattana 2014A, Lippert 2011). An exact testing method should be used with variance components estimated for each test separately and with the likelihood ratio as the test statistic to minimise the loss of power (Zhang et al. 2010, Eu-ahsunthornwattana 2014A, Lippert 2011). Currently, the most commonly used MLM programmes that offer the exact test option are EMMA, GEMMA and FAST-LMM. When used to conduct the so called 'exact' or 'full' test that determines heritable and non-heritable variance at each SNP independently, all perform extremely similarly. Essentially, as several papers and reviews

state, they should be nearly or just identical in terms of significance and P-values, but they should differ in speed with EMMA being slowest (Eu-ahsunthornwattana 2014A, Eu-ahsunthornwattana 2014B, Zhang et al. 2010). If possible, empirical 5% FDR threshold should be optimally established by permutation to take into account the sample LD-structure and estimate the precise level of genome-wide significance. Alternatively a WTCCC threshold of  $5 \times 10^{-8}$  for Caucasians can be applied or empirical threshold for the British population (approximately  $7 \times 10^{-8}$ ).

For the meta-analysis step, using GWAMA random-effect test, or preferably MANTRA (which can combine fixed effects within same-ethnicity populations and random effects across different ethnicities), would allow to combine the results from cohorts composed of different ethnicities as well as populations (Morris 2011). This could potentially increase power to detect significant associations, also the use imputed genotypes instead of the dosage format at the expense of resolution. Finally, replication should be conducted using chosen candidates.

Knowing the reciprocal relationship between EBV lytic replication, high copy number and re-infection, a new association test for latency transcript abundance could be performed with EBV load as a covariate interaction term in order to remove its effect. The analysis could also be extended onto all of the EBV lytic transcripts.

Additionally, less conservative parametric correlation tests could be applied to look at the association between EBV copy and expression in the same LCLs. This could provide more insight into the extent to which particular viral transcripts promote or inhibit lytic replication.

In a future follow-up, it may be worth revisiting the Mohr (2010) study to investigate the other significant viral QTL candidate, rs2866464 associated with the difference of LMP1

expression under normoxia and hypoxia. Similarly to the LMP1 eQTL analysis, a local association test could be conducted to evaluate potential *cis* effects of rs2866464 on the expression of the nearest 3 human genes located within 250kb on either side the candidate. In this context, it may also be interesting to see if any of the 3 genes is also characterised by significantly different expression under hypoxia and normoxia and if the difference is associated to rs2866464 genotype. Testing for the difference of expression of the 3 genes as a phenotype (instead of conducting two separate association tests for normoxia and hypoxia samples) would have the additional benefit of reducing the multiple testing correction. Also the difference of expression of LMP1 (LMP1 levels have already been quantified in the normoxia-grown samples by qPCR – Section 3.4, Chapter 3) could be used to test for local associations in chosen short fragments containing LMP1 eQTL candidates (identified in the MRCA panel and meta-analysis).

However, the most important direction in which the EBV QTL research could be forwarded should aim at replication of the key findings of the current project, in particular the copy number associations to LTBP4 and PRDM9, identified in the 1000 Genomes samples.

## Reference

1. Akay, E., et al., Interleukin 28B Gene Polymorphisms and Epstein-Barr Virus-Associated Lymphoproliferative Diseases. *Intervirol*, 2014. 57(2).
2. Akiyama, M., Identification of UACA, EXOSC9, and TauMuX2 in bovine periosteal cells by mass spectrometry and immunohistochemistry. *Anal Bioanal Chem*, 2014.
3. Albagha, O.M.E., et al., Linkage disequilibrium between polymorphisms in the human TNFRSF1B gene and their association with bone mass in perimenopausal women. *Hum Mol Genet*, 2002. 11(19): p. 2289-2295.
4. Alberts, B., J.H. Wilson, and T. Hunt, *Molecular biology of the cell*. 5th ed. 2008, New York: Garland Science. xxxiii, 1601, 90 p.
5. Altshuler, D.M., et al., An integrated map of genetic variation from 1,092 human genomes. *Nature*, 2012. 491(7422): p. 56-65.
6. Amon, W. and P.J. Farrell, Reactivation of Epstein-Barr virus from latency. *Rev Med Virol*, 2005. 15(3): p. 149-56.
7. Aouizerat, B.E., et al., GWAS for discovery and replication of genetic loci associated with sudden cardiac arrest in patients with coronary artery disease. *BMC Cardiovasc Disord*, 2011. 11: p. 29.
8. Arvey, A., I. Lo Tempera, and P.M. Lieberman, Interpreting the Epstein-Barr Virus (EBV) Epigenome Using High-Throughput Data. *Viruses-Basel*, 2013. 5(4): p. 1042-1054.
9. Ascherio, A. and K.L. Munger, Environmental risk factors for multiple sclerosis. Part I: the role of infection. *Ann Neurol*, 2007. 61(4): p. 288-99.
10. Aviel, S., et al., Degradation of the Epstein-Barr virus latent membrane protein 1 (LMP1) by the ubiquitin-proteasome pathway. Targeting via ubiquitination of the N-terminal residue. *J Biol Chem*, 2000. 275(31): p. 23491-9.
11. Azuaje, F.J., et al., Information encoded in a network of inflammation proteins predicts clinical outcome after myocardial infarction. *Bmc Medical Genomics*, 2011. 4.
12. Bahlo, M., et al., Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nature Genetics*, 2009. 41(7): p. 824-U84.
13. Baik, S.Y., et al., Identification of stathmin 1 expression induced by Epstein-Barr virus in human B lymphocytes. *Cell Prolif*, 2007. 40(2): p. 268-81.
14. Baker, C.L., et al., PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. *Genome Res*, 2014.
15. Baldanti, F., et al., High levels of Epstein-Barr virus DNA in blood of solid-organ transplant recipients and their value in predicting posttransplant lymphoproliferative disorders. *Journal of Clinical Microbiology*, 2000. 38(2): p. 613-619.
16. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, et al. (2014) Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. *PLoS Genet* 10(9): e1004663. doi:10.1371/journal.pgen.1004663
17. Baranzini, S.E., et al., Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum Mol Genet*, 2009. 18(4): p. 767-778.
18. Barrett, J.C., et al., Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, 2009. 41(6): p. 703-707.
19. Baudat, F., et al., PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science*, 2010. 327(5967): p. 836-840.

20. Bell, A.I., et al., Analysis of Epstein-Barr virus latent gene expression in endemic Burkitt's lymphoma and nasopharyngeal carcinoma tumour cells by using quantitative real-time PCR assays. *J Gen Virol*, 2006. 87(Pt 10): p. 2885-90.
21. Benders, A.A., et al., Epstein-Barr virus latent membrane protein 1 is not associated with vessel density nor with hypoxia inducible factor 1 alpha expression in nasopharyngeal carcinoma tissue. *Head Neck Pathol*, 2009. 3(4): p. 276-82.
22. Benjafield, A.V., X.L. Wang, and B.J. Morris, Tumor necrosis factor receptor 2 gene (TNFRSF1B) in genetic basis of coronary artery disease. *J Mol Med (Berl)*, 2001. 79(2-3): p. 109-15.
23. Benned-Jensen, T., et al., Ligand modulation of the Epstein-Barr virus-induced seven-transmembrane receptor EB12: identification of a potent and efficacious inverse agonist. *J Biol Chem*, 2011. 286(33): p. 29292-302.
24. Benyamin, B., P.M. Visscher, and A.F. McRae, Family-based genome-wide association studies. *Pharmacogenomics*, 2009. 10(2): p. 181-190.
25. Bharti, S., et al., Src-dependent phosphorylation of ASAP1 regulates podosomes. *Mol Cell Biol*, 2007. 27(23): p. 8271-83.
26. Bhatnagar, P., et al., Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J Hum Genet*, 2011. 56(4): p. 316-23.
27. Bieging, K.T., L.J. Anderson, and R. Longnecker, Regulation of EBV Latency by LMP2A. *Epstein-Barr Virus: Latency and Transformation*, 2010: p. 135-153.
28. Birney, E., et al., Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 2007. 447(7146): p. 799-816.
29. Bollard, C.M., C.M. Rooney, and H.E. Heslop, T-cell therapy in the treatment of post-transplant lymphoproliferative disease. *Nature Reviews Clinical Oncology*, 2012. 9(9): p. 510-519.
30. Bolstad, B.M., et al., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003. 19(2): p. 185-93.
31. Brebi, P., et al., Genome-wide methylation profiling reveals Zinc finger protein 516 () and FK-506-binding protein 6 () promoters frequently methylated in cervical neoplasia, associated with HPV status and ethnicity in a Chilean population. *Epigenetics*, 2013. 9(2).
32. Brennan, R.M., et al., Strains of Epstein-Barr virus infecting multiple sclerosis patients. *Mult Scler*, 2010. 16(6): p. 643-51.
33. Brink, A.A.T.P., et al., Presence of Epstein-Barr virus latency type III at the single cell level in post-transplantation lymphoproliferative disorders and AIDS related lymphomas. *Journal of Clinical Pathology*, 1997. 50(11): p. 911-918.
34. Buettner, M., et al., Lytic Epstein-Barr virus infection in epithelial cells but not in B-lymphocytes is dependent on Blimp1. *Journal of General Virology*, 2012. 93: p. 1059-1064.
35. Burgess, D.L., et al., Mutation of a New Sodium-Channel Gene, Scn8a, in the Mouse Mutant Motor End-Plate Disease. *Nature Genetics*, 1995. 10(4): p. 461-465.
36. Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol* 8(12): e1002822. doi:10.1371/journal.pcbi.1002822
37. Calderwood, M.A., et al., Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci U S A*, 2007. 104(18): p. 7606-11.
38. Caliskan, M., et al., The effects of EBV transformation on gene expression levels and methylation profiles. *Hum Mol Genet*, 2011. 20(8): p. 1643-52.

39. Cameron, B., et al., Prolonged illness after infectious mononucleosis is associated with altered immunity but not with increased viral load. *Journal of Infectious Diseases*, 2006. 193(5): p. 664-671.
40. Champion, E.M., et al., Repression of the Proapoptotic Cellular BIK/NBK Gene by Epstein-Barr Virus Antagonizes Transforming Growth Factor beta1-Induced B-Cell Apoptosis. *J Virol*, 2014. 88(9): p. 5001-13.
41. Campo, E., et al., The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood*, 2011. 117(19): p. 5019-5032.
42. Cantero-Recasens, G., et al., The asthma-associated ORMDL3 gene product regulates endoplasmic reticulum-mediated calcium signaling and cellular stress. *Hum Mol Genet*, 2010. 19(1): p. 111-21.
43. Carbone, A., A. Gloghini, and G. Dotti, EBV-associated lymphoproliferative disorders: classification and treatment. *Oncologist*, 2008. 13(5): p. 577-85.
44. Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001;2:91-99
45. Carreras-Sureda, A., et al., ORMDL3 modulates store-operated calcium entry and lymphocyte activation. *Hum Mol Genet*, 2013. 22(3): p. 519-30.
46. Chadburn, A., Immunodeficiency-associated lymphoid proliferations (ALPS, HIV, and KSHV/HHV8). *Seminars in Diagnostic Pathology*, 2013. 30(2): p. 113-129.
47. Chang, C.M., et al., The extent of genetic diversity of Epstein-Barr virus and its geographic and disease patterns: A need for reappraisal. *Virus Research*, 2009. 143(2): p. 209-221.
48. Chaouchi, N., et al., Characterization of transforming growth factor-beta 1 induced apoptosis in normal human B cells and lymphoma B cell lines. *Oncogene*, 1995. 11(8): p. 1615-22.
49. Chau, C.M., et al., Regulation of Epstein-Barr virus latency type by the chromatin boundary factor CTCF. *Journal of Virology*, 2006. 80(12): p. 5723-5732.
50. Chen, W.M. and G.R. Abecasis, Estimating the power of variance component linkage analysis in large pedigrees. *Genet Epidemiol*, 2006. 30(6): p. 471-484.
51. Chen, W.M. and G.R. Abecasis, Family-based association tests for genomewide association scans. *Am J Hum Genet*, 2007. 81(5): p. 913-26.
52. Cheung, C.L., et al., Meta-analysis of gene-based genome-wide association studies of bone mineral density in Chinese and European subjects. *Osteoporos Int*, 2012. 23(1): p. 131-42.
53. Cheung, V.G., et al., Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol*, 2010. 8(9).
54. Chiang, A.K., N.K. Mak, and W.T. Ng, Translational research in nasopharyngeal carcinoma. *Oral Oncol*, 2013.
55. Chisholm, C. and L. Lopez, Cutaneous infections caused by Herpesviridae: a review. *Arch Pathol Lab Med*, 2011. 135(10): p. 1357-62.
56. Choy, E.Y.W., et al., An Epstein-Barr virus-encoded microRNA targets PUMA to promote host cell survival. *Journal of Experimental Medicine*, 2008. 205(11): p. 2551-2560.
57. Cleary, S.P., et al., Identification of Driver Genes in Hepatocellular Carcinoma by Exome Sequencing. *Hepatology*, 2013. 58(5): p. 1693-1702.
58. Coffey, A.J., et al., Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nature Genetics*, 1998. 20(2): p. 129-135.
59. Cohen, J.I., Epstein-Barr virus infection. *New England Journal of Medicine*, 2000. 343(7): p. 481-492.

60. Collins, C.M., et al., The terminal repeats and latency-associated nuclear antigen of herpesvirus saimiri are essential for episomal persistence of the viral genome. *Journal of General Virology*, 2002. 83: p. 2269-2278.
61. Combadiere, B., et al., The chemokine receptor CX3CR1 controls homing and anti-viral potencies of CD8 effector-memory T lymphocytes in HIV-infected patients. *Aids*, 2003. 17(9): p. 1279-1290.
62. Cookson, W., et al., Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 2009. 10(3): p. 184-194.
63. Cooper, J.D., et al., Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics*, 2008. 40(12): p. 1399-1401.
64. Cordell, H.J. and D.G. Clayton, Genetic epidemiology 3 - Genetic association studies. *Lancet*, 2005. 366(9491): p. 1121-1131.
65. Cozen, W., et al., A protective role for early oral exposures in the etiology of young adult Hodgkin lymphoma. *Blood*, 2009. 114(19): p. 4014-4020.
66. Crackower, M.A., et al., Essential role of Fkbp6 in male fertility and homologous chromosome pairing in meiosis. *Science*, 2003. 300(5623): p. 1291-1295.
67. Crawford, D.H., Biology and disease associations of Epstein-Barr virus. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 2001. 356(1408): p. 461-473.
68. Cristina Gattazzo, A.T., Chiara Nardin, Francesca Passeri, Gregorio Barilà, Valentina Trimarco, Alberto Pavan, Samuela Carraro, Elena De March, Monica Facco, Livio Trentin, Gianpietro Semenzato and Renato Zambello,, T Large Granular Lymphocytes Leukemia (T-LGLL) and Natural Killer Chronic Lymphoproliferative Disorder (NK-CLPD): Two Diseases With a Common Etiopathogenetic Mechanism? *Blood* 2013. vol. 122 ( no. 21 ).
69. Da Silva, G.N., et al., Epstein-barr virus infection and single nucleotide Polymorphisms in the promoter region of interleukin 10 gene in patients with Hodgkin lymphoma. *Archives of Pathology & Laboratory Medicine*, 2007. 131(11): p. 1691-1696.
70. Dai, Y., et al., Screening and functional analysis of differentially expressed genes in EBV-transformed lymphoblasts. *Virology*, 2012. 9: p. 77.
71. Damaschke, N.A., et al., Epigenetic susceptibility factors for prostate cancer with aging. *Prostate*, 2013. 73(16): p. 1721-30.
72. Damerval, C., et al., Quantitative Trait Loci Underlying Gene-Product Variation - a Novel Perspective for Analyzing Regulation of Genome Expression. *Genetics*, 1994. 137(1): p. 289-301.
73. Daniel Hartman Johnson, T.M.R., Herbert Twase, Marco Ruiz and Rachna Jetly-Shridhar,, HIV-Associated Burkitt Lymphoma: Case Report and Literature Review. *Blood*, 2013. 122(21): p. 1.
74. Dawson, C.W., R.J. Port, and L.S. Young, The role of the EBV-encoded latent membrane proteins LMP1 and LMP2 in the pathogenesis of nasopharyngeal carcinoma (NPC). *Seminars in Cancer Biology*, 2012. 22(2): p. 144-153.
75. de Brouwer, A.P., H. van Bokhoven, and H. Kremer, Comparison of 12 reference genes for normalization of gene expression levels in Epstein-Barr virus-transformed lymphoblastoid cell lines and fibroblasts. *Mol Diagn Ther*, 2006. 10(3): p. 197-204.
76. De Jager, P.L., et al., Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nature Genetics*, 2009. 41(7): p. 776-82.
77. de Jonge, H.J., et al., Evidence based selection of housekeeping genes. *PLoS One*, 2007. 2(9): p. e898.

78. Dehee, A., et al., Quantification of Epstein-Barr virus load in peripheral blood of human immunodeficiency virus-infected patients using real-time PCR. *J Med Virol*, 2001. 65(3): p. 543-52.
79. Delecluse, H.J., et al., Epstein-Barr virus-associated tumours: an update for the attention of the working pathologist. *Journal of Clinical Pathology*, 2007. 60(12): p. 1358-1364.
80. Dethe, G., et al., Epidemiological Evidence for Causal Relationship between Epstein-Barr Virus and Burkitts-Lymphoma from Ugandan Prospective-Study. *Nature*, 1978. 274(5673): p. 756-761.
81. Deutsch, A.J., et al., Distinct signatures of B-cell homeostatic and activation-dependent chemokine receptors in the development and progression of extragastric MALT lymphomas. *J Pathol*, 2008. 215(4): p. 431-44.
82. Devlin, B. and K. Roeder, Genomic control for association studies. *American Journal of Human Genetics*, 1999. 65(4): p. A83-A83.
83. Dheda, K., et al., Validation of housekeeping genes for normalizing RNA expression in real-time PCR. *Biotechniques*, 2004. 37(1): p. 112-4, 116, 118-9.
84. di Renzo, L., et al., Endogenous TGF-beta contributes to the induction of the EBV lytic cycle in two Burkitt lymphoma cell lines. *Int J Cancer*, 1994. 57(6): p. 914-9.
85. Diehl, V., et al., Long-term cultivation of plasma cell leukemia cells and autologous lymphoblasts (LCL) in vitro: a comparative study. *Blut*, 1978. 36(6): p. 331-8.
86. Dillon, P.J., et al., Tousel-like kinases modulate reactivation of gammaherpesviruses from latency. *Cell Host Microbe*, 2013. 13(2): p. 204-14.
87. Dimmock, N.J., A.J. Easton, and K. Leppard, Introduction to modern virology. 6th ed. 2007, Malden, MA: Blackwell Pub. xiv, 516 p.
88. Distler, M.G., et al., Assessment of Behaviors Modeling Aspects of Schizophrenia in Csm1 Mutant Mice. *Plos One*, 2012. 7(12).
89. Dixon, A.L., et al., A genome-wide association study of global gene expression. *Nature Genetics*, 2007. 39(10): p. 1202-1207.
90. Dojcinov, S.D., et al., Age-related EBV-associated lymphoproliferative disorders in the Western population: a spectrum of reactive lymphoid hyperplasia and lymphoma. *Blood*, 2011. 117(18): p. 4726-4735.
91. Donohoe, G., et al., Neuropsychological effects of the CSMD1 genome-wide associated schizophrenia risk variant rs10503253. *Genes Brain Behav*, 2013. 12(2): p. 203-9.
92. Drutel, G., et al., ARNT2, a transcription factor for brain neuron survival? *European Journal of Neuroscience*, 1999. 11(5): p. 1545-1553.
93. Du, C., et al., CDH4 as a novel putative tumor suppressor gene epigenetically silenced by promoter hypermethylation in nasopharyngeal carcinoma. *Cancer Lett*, 2011. 309(1): p. 54-61.
94. Dubinsky M, Brant SR, Silverberg M, et al. with Rivas MA, Beaudoin M, Gardet A et al., NIDDK IBD Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics* 2011;43(11):1066-1073. doi:10.1038/ng.952.
95. Dudziak, D., et al., Latent membrane protein 1 of Epstein-Barr virus induces CD83 by the NF-kappaB signaling pathway. *Journal of Virology*, 2003. 77(15): p. 8290-8.
96. Dunham, I., et al., An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012. 489(7414): p. 57-74.
97. Efremov, R., et al., Human chemokine receptors CCR5, CCR3 and CCR2B share common polarity motif in the first extracellular loop with other human G-protein

- coupled receptors - Implications for HIV-1 coreceptor function. *European Journal of Biochemistry*, 1999. 263(3): p. 746-756.
98. Ehlers, C., et al., Post-transcriptional regulation of CD83 expression by AUF1 proteins. *Nucleic Acids Res*, 2013. 41(1): p. 206-219.
  99. Elias A. Rahal, L.F., Dalal F. Jaber, Aline Kherlopian, Bassem Yamout and Alexander M. Abdelnoor, The Effect of HLA, IL2RA and IL7RA Alleles on the Risk of Multiple Sclerosis in HHV-6 Infected and Uninfected Lebanese Subjects. *Research in Immunology*, 2014. Vol. 2014: p. 7 pages.
  100. Elliott, J., et al., Variable methylation of the Epstein-Barr virus Wp EBNA gene promoter in B-lymphoblastoid cell lines. *J Virol*, 2004. 78(24): p. 14062-5.
  101. email, S.T.a.P.J.F., Epstein-Barr Virus Sequence Variation-Biology and Disease. *Pathogens* 2012. 1(2).
  102. Engel, P., M.J. Eck, and C. Terhorst, The SAP and SLAM families in immune responses and X-linked lymphoproliferative disease. *Nature Reviews Immunology*, 2003. 3(10): p. 813-821.
  103. Eu-ahsunthornwattana, Jakris, et al. "Comparison of methods to account for relatedness in genome-wide association studies with family-based data." *PLoS genetics* 10.7 (2014): e1004445.
  104. Eu-ahsunthornwattana, Jakris, Richard AJ Howey, and Heather J. Cordell. "Accounting for relatedness in family-based association studies: application to Genetic Analysis Workshop 18 data." *BMC proceedings*. Vol. 8. No. Suppl 1. BioMed Central Ltd, 2014.
  105. Evans, T.J., P.J. Farrell, and S. Swaminathan, Molecular genetic analysis of Epstein-Barr virus Cp promoter function. *Journal of Virology*, 1996. 70(3): p. 1695-1705.
  106. Evdokimov, K., et al., Proteolytic cleavage of LEDA-1/PIANP by furin-like proprotein convertases precedes its plasma membrane localization. *Biochemical and Biophysical Research Communications*, 2013. 434(1): p. 22-27.
  107. Farrell, P.J., et al., Direct demonstration of persistent Epstein-Barr virus gene expression in peripheral blood of infected common marmosets and analysis of virus-infected tissues in vivo. *J Gen Virol*, 1997. 78 ( Pt 6): p. 1417-24.
  108. Faure, S., et al., Deleterious genetic influence of CX3CR1 genotypes on HIV-1 disease progression. *J AIDS-Journal of Acquired Immune Deficiency Syndromes*, 2003. 32(3): p. 335-337.
  109. Feingold, E.A., et al., The ENCODE (ENCyclopedia of DNA elements) Project. *Science*, 2004. 306(5696): p. 636-640.
  110. Ferrera, L., A. Caputo, and L.J. Galletta, TMEM16A protein: a new identity for Ca(2+)-dependent Cl(-) channels. *Physiology (Bethesda)*, 2010. 25(6): p. 357-63.
  111. Filipovich, A.H., et al., X-linked lymphoproliferative syndromes: brothers or distant cousins? *Blood*, 2010. 116(18): p. 3398-3408.
  112. Foster, A.E., et al., Antitumor activity of EBV-specific T lymphocytes transduced with a dominant negative TGF-beta receptor. *Journal of Immunotherapy*, 2008. 31(5): p. 500-5.
  113. Fox, C.P., C. Shannon-Lowe, and M. Rowe, Deciphering the role of Epstein-Barr virus in the pathogenesis of T and NK cell lymphoproliferations. *Herpesviridae*, 2011. 2: p. 8.
  114. Franke, A., et al., Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, 2010. 42(12): p. 1118-+.
  115. Frappier, L., EBNA1 in Viral DNA Replication and Persistence. *Epstein-Barr Virus: Latency and Transformation*, 2010: p. 37-59.

116. Frappier, L., Promyelocytic leukemia nuclear body disruption as a treatment for EBV-associated nasopharyngeal carcinoma? *Future Virology*, 2012. 7(1): p. 1-3.
117. Frappier, L., Viral disruption of promyelocytic leukemia (PML) nuclear bodies by hijacking host PML regulators. *Virulence*, 2011. 2(1): p. 58-62.
118. Freedman, M.L., et al., Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genetics*, 2011. 43(6): p. 513-8.
119. Freson, K., et al., Silencing of RhoA nucleotide exchange factor, ARHGEF3, reveals its role in thrombopoiesis and iron uptake. *Journal of Thrombosis and Haemostasis*, 2011. 9: p. 33-33.
120. Freudenberg, J., et al., Genome-Wide Association Study of Rheumatoid Arthritis in Koreans. *Arthritis and Rheumatism*, 2011. 63(4): p. 884-893.
121. Fujikane, T., et al., Genomic screening for genes upregulated by demethylation revealed novel targets of epigenetic silencing in breast cancer. *Breast Cancer Research and Treatment*, 2010. 122(3): p. 699-710.
122. Fukayama, M., et al., Epstein-Barr virus-associated gastric carcinoma and Epstein-Barr virus infection of the stomach. *Lab Invest*, 1994. 71(1): p. 73-81.
123. Fuller, C.M., Time for TMEM? *J Physiol*, 2012. 590(Pt 23): p. 5931-2.
124. Gallego-Paez, L.M., et al., Smc5/6-mediated regulation of replication progression contributes to chromosome assembly during mitosis in human cells. *Mol Biol Cell*, 2014. 25(2): p. 302-17.
125. Gehman, L.T., et al., The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nature Genetics*, 2011. 43(7): p. 706-U133.
126. George, L.C., M. Rowe, and C.P. Fox, Epstein-barr virus and the pathogenesis of T and NK lymphoma: a mystery unsolved. *Curr Hematol Malig Rep*, 2012. 7(4): p. 276-84.
127. Gibbs, R.A., et al., The International HapMap Project. *Nature*, 2003. 426(6968): p. 789-796.
128. Gibson, S.E. and E.D. Hsi, Epstein-Barr virus-positive B-cell lymphoma of the elderly at a United States tertiary medical center: an uncommon aggressive lymphoma with a nongerminal center B-cell phenotype. *Human Pathology*, 2009. 40(5): p. 653-661.
129. Gilad, Y., S.A. Rifkin, and J.K. Pritchard, Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*, 2008. 24(8): p. 408-415.
130. Gilbert, M.R., R. Ruda, and R. Soffiatti, Ependymomas. *Primary Central Nervous System Tumors: Pathogenesis and Therapy*, 2011: p. 249-262.
131. Goldman, A.S., et al., Immunodeficiency due to a unique protracted developmental delay in the B-cell lineage. *Clinical and Diagnostic Laboratory Immunology*, 1999. 6(2): p. 161-167.
132. Greif, P.A., et al., Identification of Recurring Tumor-Specific Somatic Mutations in Acute Myeloid Leukaemia by Transcriptome Sequencing. *Annals of Hematology*, 2011. 90: p. S9-S9.
133. Grogg, K.L., R.F. Miller, and A. Dogan, HIV infection and lymphoma. *Journal of Clinical Pathology*, 2007. 60(12): p. 1365-1372.
134. Hadinoto, V., et al., On the dynamics of acute EBV infection and the pathogenesis of infectious mononucleosis. *Blood*, 2008. 111(3): p. 1420-1427.
135. Hahn, A.S. and R.C. Desrosiers, Rhesus monkey rhadinovirus uses eph family receptors for entry into B cells and endothelial cells but not fibroblasts. *PLoS Pathog*, 2013. 9(5): p. e1003360.
136. Hahn, A.S., et al., The ephrin receptor tyrosine kinase A2 is a cellular receptor for Kaposi's sarcoma-associated herpesvirus. *Nature Medicine*, 2012. 18(6): p. 961-+.

137. Hankinson, O., The role of the aryl hydrocarbon receptor nuclear translocator protein in aryl hydrocarbon receptor action. *Trends Endocrinol Metab*, 1994. 5(6): p. 240-4.
138. Hansell, N.K., et al., Linkage analysis of alcohol dependence symptoms in the community. *Alcohol Clin Exp Res*, 2010. 34(1): p. 158-63.
139. Hardie, D.R., Human gamma-herpesviruses: A review of 2 divergent paths to oncogenesis. *Transfusion and Apheresis Science*, 2010. 42(2): p. 177-183.
140. Hartford, C.M., et al., Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. *Blood*, 2009. 113(10): p. 2145-53.
141. Hasbold, J., et al., Quantitative analysis of lymphocyte differentiation and proliferation in vitro using carboxyfluorescein diacetate succinimidyl ester. *Immunology and Cell Biology*, 1999. 77(6): p. 516-522.
142. Havik, B., et al., The Complement Control-Related Genes CSMD1 and CSMD2 Associate to Schizophrenia. *Biological Psychiatry*, 2011. 70(1): p. 35-42.
143. Hawinkels, L.J. and P. Ten Dijke, Exploring anti-TGF-beta therapies in cancer and fibrosis. *Growth Factors*, 2011. 29(4): p. 140-52.
144. He, B., X.G. Qiao, and A. Cerutti, CpG DNA induces IgG class switch DNA recombination by activating human B cells through an innate pathway that requires TLR9 and cooperates with IL-10. *Journal of Immunology*, 2004. 173(7): p. 4479-4491.
145. Heo, J.I., J.H. Cho, and J.R. Kim, HJURP regulates cellular senescence in human fibroblasts and endothelial cells via a p53-dependent pathway. *J Gerontol A Biol Sci Med Sci*, 2013. 68(8): p. 914-25.
146. Hernando, H., et al., Epstein-Barr virus-mediated transformation of B cells induces global chromatin changes independent to the acquisition of proliferation. *Nucleic Acids Res*, 2014. 42(1): p. 249-63.
147. Hildesheim, A. and C.P. Wang, Genetic predisposition factors and nasopharyngeal carcinoma risk: a review of epidemiological association studies, 2000-2011: Rosetta Stone for NPC: genetics, viral infection, and other environmental factors. *Seminars in Cancer Biology*, 2012. 22(2): p. 107-16.
148. Hiraki, A., et al., Genetics of Epstein-Barr virus infection. *Biomed Pharmacother*, 2001. 55(7): p. 369-72.
149. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6: 95-108.
150. Hjalgrim, H., et al., HLA-A alleles and infectious mononucleosis suggest a critical role for cytotoxic T-cell response in EBV-related Hodgkin lymphoma. *Proc Natl Acad Sci U S A*, 2010. 107(14): p. 6400-5.
151. Hochberg, D., et al., Acute infection with Epstein-Barr virus targets and overwhelms the peripheral memory B-cell compartment with resting, latently infected cells. *Journal of Virology*, 2004. 78(10): p. 5194-204.
152. Hohaus, S., et al., The viral load of Epstein-Barr virus (EBV) DNA in peripheral blood predicts for biological and clinical characteristics in Hodgkin lymphoma. *Clin Cancer Res*, 2011. 17(9): p. 2885-92.
153. Holm, K., et al., SNPexp - A web tool for calculating and visualizing correlation between HapMap genotypes and gene expression levels. *BMC Bioinformatics*, 2010. 11: p. 600.
154. Hopwood, P. and D.H. Crawford, The role of EBV in post-transplant malignancies: a review. *Journal of Clinical Pathology*, 2000. 53(4): p. 248-254.
155. Horuk, R., et al., The CC chemokine I-309 inhibits CCR8-dependent infection by diverse HIV-1 strains. *Journal of Biological Chemistry*, 1998. 273(1): p. 386-391.

156. Houldcroft CJ, Petrova V, Liu JZ, Frampton D, Anderson CA, et al. (2014) Host Genetic Variants and Gene Expression Patterns Associated with Epstein-Barr Virus Copy Number in Lymphoblastoid Cell Lines. *PLoS ONE* 9(10): e108384. doi: 10.1371/journal.pone.0108384
157. Houldsworth, A., et al., Polymorphisms in the IL-12B gene and outcome of HCV infection. *Journal of Interferon and Cytokine Research*, 2005. 25(5): p. 271-276.
158. Hsu, J.L. and S.L. Glaser, Epstein-barr virus-associated malignancies: epidemiologic patterns and etiologic implications. *Crit Rev Oncol Hematol*, 2000. 34(1): p. 27-53.
159. Hsu, K.J. and S.E. Turvey, Functional analysis of the impact of ORMDL3 expression on inflammation and activation of the unfolded protein response in human airway epithelial cells. *Allergy Asthma and Clinical Immunology*, 2013. 9.
160. Huang, X., et al., Multiple HLA class I and II associations in classical Hodgkin lymphoma and EBV status defined subgroups (Retraction of vol 118, pg 5211, 2011). *Blood*, 2012. 119(14): p. 3370-3370.
161. Huang, Y., et al., Gene expression profiling identifies emerging oncogenic pathways operating in extranodal NK/T-cell lymphoma, nasal type. *Blood*, 2010. 115(6): p. 1226-37.
162. Huggett, J., et al., Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun*, 2005. 6(4): p. 279-84.
163. Hussin, J., et al., Rare allelic forms of PRDM9 associated with childhood leukemogenesis. *Genome Res*, 2013. 23(3): p. 419-30.
164. Ibrahim, H.A. and K.N. Naresh, Posttransplant lymphoproliferative disorders. *Adv Hematol*, 2012. 2012: p. 230173.
165. Ikegaya, H., et al., Forensic application of Epstein-Barr virus genotype: correlation between viral genotype and geographical area. *J Virol Methods*, 2008. 147(1): p. 78-85.
166. Inman, G.J. and M.J. Allday, Resistance to TGF-beta1 correlates with a reduction of TGF-beta type II receptor expression in Burkitt's lymphoma and Epstein-Barr virus-transformed B lymphoblastoid cell lines. *J Gen Virol*, 2000. 81(Pt 6): p. 1567-78.
167. Inokawa, Y., et al., Dynamin 3: a new candidate tumor suppressor gene in hepatocellular carcinoma detected by triple combination array analysis. *Onco Targets Ther*, 2013. 6: p. 1417-24.
168. Irizarry, R.A., et al., Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, 2003. 31(4): p. e15.
169. Izumi, K.M., The EBV Latent Membrane Protein 1 Oncoprotein. *Epstein-Barr Virus: Latency and Transformation*, 2010: p. 117-134.
170. Jaccard, A., et al., L-asparaginase-based treatment of 15 western patients with extranodal NK/T-cell lymphoma and leukemia and a review of the literature. *Ann Oncol*, 2009. 20(1): p. 110-6.
171. Jacobson, C.A. and A.S. LaCasce, Lymphoma: Risk and Response After Solid Organ Transplant. *Oncology-New York*, 2010. 24(10): p. 936-944.
172. Jang, K.L., et al., Up-regulation of beta-catenin by a viral oncogene correlates with inhibition of the seven in absentia homolog 1 in B lymphoma cells. *Proc Natl Acad Sci U S A*, 2005. 102(51): p. 18431-6.
173. Jia, W.H. and H.D. Qin, Non-viral environmental risk factors for nasopharyngeal carcinoma: a systematic review. *Seminars in Cancer Biology*, 2012. 22(2): p. 117-26.
174. Jinno, A., et al., Identification of the chemokine receptor TER1/CCR8 expressed in brain-derived cells and T cells as a new coreceptor for HIV-1 infection. *Biochemical and Biophysical Research Communications*, 1998. 243(2): p. 497-502.

175. Jochum, S., et al., RNAs in Epstein-Barr virions control early steps of infection. *Proceedings of the National Academy of Sciences of the United States of America*, 2012. 109(21): p. E1396-E1404.
176. Jung, S.K., et al., Crystal structure of human slingshot phosphatase 2. *Proteins*, 2007. 68(1): p. 408-12.
177. Kaarvatn, M.H., et al., Single Nucleotide Polymorphism in the Interleukin 12B Gene is Associated with Risk for Breast Cancer Development. *Scandinavian Journal of Immunology*, 2012. 76(3): p. 329-335.
178. Kaji, N., et al., Cell cycle-associated changes in Slingshot phosphatase activity and roles in cytokinesis in animal cells. *J Biol Chem*, 2003. 278(35): p. 33450-5.
179. Kampert, M.M.D., *Statistical Analysis in Genome-Wide Association Studies on GenoType-Imputed Family Data: A Research Strategy to Compare Various Toolsets, in Mathematics*. 2011, Leiden University: Leiden.
180. Kanarek, N. and Y. Ben-Neriah, Regulation of NF-kappaB by ubiquitination and degradation of the IkappaBs. *Immunol Rev*, 2012. 246(1): p. 77-94.
181. Kasahara, Y., et al., Differential cellular targets of Epstein-Barr virus (EBV) infection between acute EBV-associated hemophagocytic lymphohistiocytosis and chronic active EBV infection. *Blood*, 2001. 98(6): p. 1882-1888.
182. Kaser A, Zeissig S, Blumberg RS (2010) Inflammatory bowel disease. *Annu Rev Immunol* 28: 573-621 doi:10.1146/annurev-immunol-030409-101225
183. Kasowski, M., et al., Variation in Transcription Factor Binding Among Humans. *Science*, 2010. 328(5975): p. 232-235.
184. Kataoka, M., et al., Aberration of p53 and DCC in gastric and colorectal cancer. *Oncol Rep*, 2000. 7(1): p. 99-103.
185. Kato, H., et al., Gene Expression Profiling of Age-Related Epstein-Barr Virus (EBV)-Associated B-Cell Lymphoproliferative Disorder Uncovers Alterations in Immune and Inflammatory Genes: Possible Implications for Pathogenesis. *Blood*, 2011. 118(21): p. 1474-1474.
186. Kawasaki, A., et al., Analysis on the association of human BLYS (BAFF, TNFSF13B) polymorphisms with systemic lupus erythematosus and rheumatoid arthritis. *Genes Immun*, 2002. 3(7): p. 424-9.
187. Kehrl, J.H., et al., Production of transforming growth factor beta by human T lymphocytes and its potential role in the regulation of T cell growth. *Journal of Experimental Medicine*, 1986. 163(5): p. 1037-50.
188. Kelly, G.L., et al., Three restricted forms of Epstein-Barr virus latency counteracting apoptosis in c-myc-expressing Burkitt lymphoma cells. *Proc Natl Acad Sci U S A*, 2006. 103(40): p. 14935-14940.
189. Kempkes, B., EBNA-2 in Transcription Activation of Viral and Cellular Genes. *Epstein-Barr Virus: Latency and Transformation*, 2010: p. 61-80.
190. Kempkes, B., EBNA-2 in Transcription Activation of Viral and Cellular Genes. *Epstein-Barr Virus: Latency and Transformation*, 2010: p. 61-80.
191. Kenney, S.C. and J.E. Mertz, Regulation of the latent-lytic switch in Epstein-Barr virus. *Seminars in Cancer Biology*, 2014.
192. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. *Nature* 2011;474(7351):307-317. doi:10.1038/nature10209.
193. Kieff, E., E. Johannsen, and M.A. Calderwood, Latent Epstein-Barr Virus Infections. *Epstein-Barr Virus: Latency and Transformation*, 2010: p. 1-24.
194. Kieff, E., E. Johannsen, and M.A. Calderwood, Latent Epstein-Barr Virus Infections. *Epstein-Barr Virus: Latency and Transformation*, 2010: p. 1-24.

195. Kim, Y.D., et al., NSrp70 is a novel nuclear speckle-related protein that modulates alternative pre-mRNA splicing in vivo. *Nucleic Acids Res*, 2011. 39(10): p. 4300-14.
196. Kimura, H., et al., Clinical and virologic characteristics of chronic active Epstein-Barr virus infection. *Blood*, 2001. 98(2): p. 280-6.
197. Kitagawa, N., et al., Expression of seven-in-absentia homologue 1 and hypoxia-inducible factor 1 alpha: novel prognostic factors of nasopharyngeal carcinoma. *Cancer Lett*, 2013. 331(1): p. 52-7.
198. Kligys, K., et al., The slingshot family of phosphatases mediates Rac1 regulation of cofilin phosphorylation, laminin-332 organization, and motility behavior of keratinocytes. *J Biol Chem*, 2007. 282(44): p. 32520-8.
199. Knight, J.C., *Human genetic diversity : functional consequences for health and disease*. 2009, Oxford ; New York: Oxford University Press. xix, 480 p.
200. Kondo, K., et al., Inhibition of HIF is necessary for tumor suppression by the von Hippel-Lindau protein. *Cancer Cell*, 2002. 1(3): p. 237-246.
201. Kraus, D.M., et al., CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. *Journal of Immunology*, 2006. 176(7): p. 4419-4430.
202. Kruglyak, Leonid. "The use of a genetic map of biallelic markers in linkage studies." *Nature genetics* 17.1 (1997): 21-24.
203. Kuo, T.T., L.Y. Shih, and N.M. Tsang, Nasal NK/T cell lymphoma in Taiwan: A clinicopathologic study of 22 cases, with analysis of histologic subtypes, Epstein-Barr virus LMP-1 gene association, and treatment modalities. *International Journal of Surgical Pathology*, 2004. 12(4): p. 375-387.
204. Kuppers, R., The biology of Hodgkin's lymphoma. *Nature Reviews Cancer*, 2009. 9(1): p. 15-27.
205. Lagneaux, L., et al., Heterogenous response of B lymphocytes to transforming growth factor-beta in B-cell chronic lymphocytic leukaemia: Correlation with the expression of TGF-beta receptors. *British Journal of Haematology*, 1997. 97(3): p. 612-620.
206. Lango Allen, H., et al., Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 2010. 467(7317): p. 832-8.
207. Lecat, A., et al., The c-Jun N-terminal Kinase (JNK)-binding Protein (JNKBP1) Acts as a Negative Regulator of NOD2 Protein Signaling by Inhibiting Its Oligomerization Process. *Journal of Biological Chemistry*, 2012. 287(35): p. 29213-29226.
208. Lee, S., et al., CCR8 on human thymocytes functions as a human immunodeficiency virus type 1 coreceptor. *Journal of Virology*, 2000. 74(15): p. 6946-6952.
209. Leonard, S., et al., Epigenetic and Transcriptional Changes Which Follow Epstein-Barr Virus Infection of Germinal Center B Cells and Their Relevance to the Pathogenesis of Hodgkin's Lymphoma. *Journal of Virology*, 2011. 85(18): p. 9568-9577.
210. Letterio, J. J. (2005). TGF- $\beta$  signaling in T cells: roles in lymphoid and epithelial neoplasia. *Oncogene* 24: 5701-5712.
211. Li, G.H., A.W. Kung, and Q.Y. Huang, Common variants in FLNB/CRTAP, not ARHGEF3 at 3p, are associated with osteoporosis in southern Chinese women. *Osteoporos Int*, 2010. 21(6): p. 1009-20.
212. Liang, L.M., et al., A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Research*, 2013. 23(4): p. 716-726.
213. Lieberman, P.M., Keeping it quiet: chromatin control of gammaherpesvirus latency. *Nature Reviews Microbiology*, 2013. 11(12): p. 863-875.

214. Lin, N., et al., Deletion or epigenetic silencing of AJAP1 on 1p36 in glioblastoma. *Mol Cancer Res*, 2012. 10(2): p. 208-17.
215. Lindsey, J.W., et al., Quantitative PCR for Epstein-Barr virus DNA and RNA in multiple sclerosis. *Mult Scler*, 2009. 15(2): p. 153-8.
216. Ling, P.D., et al., EBNA-2 upregulation of Epstein-Barr virus latency promoters and the cellular CD23 promoter utilizes a common targeting intermediate, CBF1. *J Virol*, 1994. 68(9): p. 5375-83.
217. Linka, R.M., et al., Loss-of-function mutations within the IL-2 inducible kinase ITK in patients with EBV-associated lymphoproliferative diseases. *Leukemia*, 2012. 26(5): p. 963-971.
218. Lippert, Christoph, et al. "FaST linear mixed models for genome-wide association studies." *Nature Methods* 8.10 (2011): 833-835.
219. Liu, J.P., et al., Mechanisms of cell immortalization mediated by EB viral activation of telomerase in nasopharyngeal carcinoma. *Cell Research*, 2006. 16(10): p. 809-817.
220. Llanora, G.V., C.M. Tay, and H.P. van Bever, Gianotti-Crosti syndrome: case report of a pruritic acral exanthema in a child. *Asia Pac Allergy*, 2012. 2(3): p. 223-6.
221. Lo, A.W. and W.K. Newey, A Large-Sample Chow Test for the Linear Simultaneous Equation. *Economics Letters*, 1985. 18(4): p. 351-353.
222. Long, X. and J.M. Miano, Transforming growth factor-beta1 (TGF-beta1) utilizes distinct pathways for the transcriptional activation of microRNA 143/145 in human coronary artery smooth muscle cells. *J Biol Chem*, 2011. 286(34): p. 30119-29.
223. Lonsdale, J., et al., The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 2013. 45(6): p. 580-585.
224. Lord, J.C., et al., Evaluation of quantitative PCR reference genes for gene expression studies in *Tribolium castaneum* after fungal challenge. *J Microbiol Methods*, 2010. 80(2): p. 219-21.
225. Lossius, A., et al., Epstein-Barr Virus in Systemic Lupus Erythematosus, Rheumatoid Arthritis and Multiple Sclerosis-Association and Causation. *Viruses-Basel*, 2012. 4(12): p. 3701-3730.
226. Louis, C.U., et al., Adoptive Transfer of EBV-specific T Cells Results in Sustained Clinical Responses in Patients With Locoregional Nasopharyngeal Carcinoma. *Journal of Immunotherapy*, 2010. 33(9): p. 983-990.
227. Lucas, R.M., et al., Epstein-Barr virus and multiple sclerosis. *Journal of Neurology Neurosurgery and Psychiatry*, 2011. 82(10): p. 1142-1148.
228. Luenemann, J.D., Epstein-Barr virus in multiple sclerosis A continuing conundrum. *Neurology*, 2012. 78(1): p. 11-12.
229. Lunemann, J.D. and C. Munz, EBV in MS: guilty by association? *Trends Immunol*, 2009. 30(6): p. 243-8.
230. Lutichau, H.R., et al., A highly selective CC chemokine receptor (CCR)8 antagonist encoded by the poxvirus *molluscum contagiosum*. *Journal of Experimental Medicine*, 2000. 191(1): p. 171-179.
231. Mackay, T.F., E.A. Stone, and J.F. Ayroles, The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 2009. 10(8): p. 565-77.
232. MacMahon, E.M., et al., Epstein-Barr virus in AIDS-related primary central nervous system lymphoma. *Lancet*, 1991. 338(8773): p. 969-73.
233. Macsween, K.F. and D.H. Crawford, Epstein-Barr virus-recent advances. *Lancet Infect Dis*, 2003. 3(3): p. 131-40.
234. Maeda, A., T. Sato, and H. Wakiguchi, [Epidemiology of Epstein-Barr virus (EBV) infection and EBV-associated diseases]. *Nihon Rinsho*, 2006. 64 Suppl 3: p. 609-12.

235. Magi, R. and A.P. Morris, GWAMA: software for genome-wide association meta-analysis. *Bmc Bioinformatics*, 2010. 11.
236. Majewski, I.J., et al., An alpha-E-catenin (CTNNA1) mutation in hereditary diffuse gastric cancer. *J Pathol*, 2013. 229(4): p. 621-9.
237. Makoshi, T., et al., Detection of epstein-barr virus in nasal T-cell lymphoma. *Acta Otolaryngol Suppl*, 2002(547): p. 46-9.
238. Maltepe, E., et al., The role of ARNT2 in tumor angiogenesis and the neural response to hypoxia. *Biochemical and Biophysical Research Communications*, 2000. 273(1): p. 231-238.
239. Mancao, C., et al., Rescue of "crippled" germinal center B cells from apoptosis by Epstein-Barr virus. *Blood*, 2005. 106(13): p. 4339-4344.
240. Marquitz, A.R. and N. Raab-Traub, The role of miRNAs and EBV BARTs in NPC. *Seminars in Cancer Biology*, 2012. 22(2): p. 166-172.
241. Marsh, R.A., et al., XIAP deficiency: a unique primary immunodeficiency best classified as X-linked familial hemophagocytic lymphohistiocytosis and not as X-linked lymphoproliferative disease. *Blood*, 2010. 116(7): p. 1079-1082.
242. Martin, N.W., et al., Educational Attainment: A Genome Wide Association Study in 9538 Australians. *PLoS One*, 2011. 6(6).
243. McClellan, M.J., et al., Downregulation of integrin receptor-signaling genes by Epstein-Barr virus EBNA 3C via promoter-proximal and -distal binding elements. *J Virol*, 2012. 86(9): p. 5165-78.
244. McFadden, K. and M.A. Luftig, Interplay between DNA tumor viruses and the host DNA damage response. *Curr Top Microbiol Immunol*, 2013. 371: p. 229-57.
245. McHeyzer-Williams, M., et al., Molecular programming of B cell memory. *Nat Rev Immunol*, 2012. 12(1): p. 24-34.
246. McIntosh, B.E., J.B. Hogenesch, and C.A. Bradfield, Mammalian Per-Arnt-Sim Proteins in Environmental Adaptation. *Annual Review of Physiology*, 2010. 72: p. 625-645.
247. Medrano, L.M., et al., Role of TNFRSF1B polymorphisms in the response of Crohn's disease patients to infliximab. *Human Immunology*, 2014. 75(1): p. 71-75.
248. Meier, U.C., et al., Translational Mini-Review Series on B cell subsets in disease. B cells in multiple sclerosis: drivers of disease pathogenesis and Trojan horse for Epstein-Barr virus entry to the central nervous system? *Clinical and Experimental Immunology*, 2012. 167(1): p. 1-6.
249. Mero, I.L., et al., Oligoclonal Band Status in Scandinavian Multiple Sclerosis Patients Is Associated with Specific Genetic Risk Alleles. *Plos One*, 2013. 8(3).
250. Miller, M., et al., ORMDL3 Transgenic Mice Have Increased Airway Remodeling and Airway Responsiveness Characteristic of Asthma. *J Immunol*, 2014. 192(8): p. 3475-87.
251. Moffatt, M.F., et al., Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, 2007. 448(7152): p. 470-U5.
252. Molyneux, E.M., et al., Burkitt's lymphoma. *Lancet*, 2012. 379(9822): p. 1234-1244.
253. Montanari, F., et al., Monomorphic T-cell post-transplant lymphoproliferative disorders exhibit markedly inferior outcomes compared to monomorphic B-cell post-transplant lymphoproliferative disorders. *Leukemia & Lymphoma*, 2010. 51(9): p. 1761-1764.
254. Montgomery, S.B., et al., Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 2010. 464(7289): p. 773-U151.
255. Morinet, F., et al., Oxygen tension level and human viral infections. *Virology*, 2013. 444(1-2): p. 31-36.

256. Morris, A.P., Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol*, 2011. 35(8): p. 809-22.
257. Motsch, N., et al., MicroRNA profiling of Epstein-Barr virus-associated NK/T-cell lymphomas by deep sequencing. *PLoS One*, 2012. 7(8): p. e42193.
258. Mueller, S., et al., Senile EBV-associated B-cell lymphoproliferative disorder of prepatellar bursa in an elderly patient with multifocal urate arthropathy. *Hematol Oncol*, 2007. 25(3): p. 140-2.
259. Munch, M., et al., A single subtype of Epstein-Barr virus in members of multiple sclerosis clusters. *Acta Neurol Scand*, 1998. 98(6): p. 395-9.
260. Munoz, J.J., et al., Expression and function of the Eph A receptors and their ligands ephrins A in the rat thymus. *Journal of Immunology*, 2002. 169(1): p. 177-184.
261. Munro, K.M., et al., EphA4 Receptor Tyrosine Kinase Is a Modulator of Onset and Disease Severity of Experimental Autoimmune Encephalomyelitis (EAE). *PLoS One*, 2013. 8(2).
262. Munro, K.M., V.M. Perreau, and A.M. Turnley, Differential Gene Expression in the EphA4 Knockout Spinal Cord and Analysis of the Inflammatory Response Following Spinal Cord Injury. *PLoS One*, 2012. 7(5).
263. Murata, T. and T. Tsurumi, Switching of EBV cycles between latent and lytic states. *Rev Med Virol*, 2014. 24(3): p. 142-53.
264. Mutalima, N., et al., Associations between Burkitt Lymphoma among Children in Malawi and Infection with HIV, EBV and Malaria: Results from a Case-Control Study. *PLoS One*, 2008. 3(6).
265. Muti, G., et al., Epstein-Barr virus (EBV) load and interleukin-10 in EBV-positive and EBV-negative post-transplant lymphoproliferative disorders. *Br J Haematol*, 2003. 122(6): p. 927-33.
266. Muti, G., et al., Significance of Epstein-Barr virus (EBV) load and interleukin-10 in post-transplant lymphoproliferative disorders. *Leuk Lymphoma*, 2005. 46(10): p. 1397-407.
267. Myers S, Bowden R, Tumian A, et al. (2010) Drive Against Hotspot Motifs in Primates Implicates the PRDM9 gene in Meiotic Recombination: We provide evidence that a rapidly evolving gene is involved in determining recombination hotspot locations in humans. *Science* (New York, N.Y.);327(5967):10.1126/science.1182363. doi:10.1126/science.1182363.
268. Nakai, H., et al., Host factors associated with the kinetics of Epstein-Barr virus DNA load in patients with primary Epstein-Barr virus infection. *Microbiology and Immunology*, 2012. 56(2): p. 93-98.
269. Nakayama, T., et al., Human B cells immortalized with Epstein-Barr virus upregulate CCR6 and CCR10 and downregulate CXCR4 and CXCR5. *Journal of Virology*, 2002. 76(6): p. 3072-7.
270. N'Diaye, A., et al., Identification, replication, and fine-mapping of Loci associated with adult height in individuals of african ancestry. *PLoS Genet*, 2011. 7(10): p. e1002298.
271. Nelson, G., et al., NF- $\kappa$ B signalling is inhibited by glucocorticoid receptor and STAT6 via distinct mechanisms. *Journal of Cell Science*, 2003. 116(12): p. 2495-2503.
272. Neurath M. F. (2014). Cytokines in inflammatory bowel disease. *Nat. Rev. Immunol.* 14 329-342 10.1038/nri3661
273. Newey, A.W.L.W.K., A Residuals-Based Wald Test for the Linear Simultaneous Equation, in Rodney L. White Center for Financial Research Working Papers, W.S.R.L.W.C.f.F.R. i, Editor., Wharton School Rodney L. White Center for Financial Research

274. Newton-Cheh, C., et al., Genome-wide association study of electrocardiographic and heart rate variability traits: the Framingham Heart Study. *BMC Med Genet*, 2007. 8 Suppl 1: p. S7.
275. Niens, M., et al., HLA-A\*02 is associated with a reduced risk and HLA-A\*01 with an increased risk of developing EBV Hodgkin lymphoma. *Blood*, 2007. 110(9): p. 3310-3315.
276. Niu, N.F., et al., Radiation pharmacogenomics: A genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines. *Genome Research*, 2010. 20(11): p. 1482-1492.
277. Novak, A.J., et al., Genetic Variation in B-Cell-Activating Factor Is Associated with an Increased Risk of Developing B-Cell Non-Hodgkin Lymphoma. *Cancer Res*, 2009. 69(10): p. 4217-4224.
278. Odumade, O.A., K.A. Hogquist, and H.H. Balfour, Progress and Problems in Understanding and Managing Primary Epstein-Barr Virus Infections. *Clinical Microbiology Reviews*, 2011. 24(1): p. 193-+.
279. Orozco, G., et al., Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. *Arthritis Rheumatol*, 2014. 66(1): p. 24-30.
280. Oxelius, V.A. and J.P. Pandey, Human immunoglobulin constant heavy G chain (IGHG) (Fc gamma) (GM) genes, defining innate variants of IgG molecules and B cells, have impact on disease and therapy. *Clinical Immunology*, 2013. 149(3): p. 475-486.
281. Oyama, T., et al., Senile EBV plus B-cell Lymphoproliferative disorders - A clinicopathologic study of 22 patients. *American Journal of Surgical Pathology*, 2003. 27(1): p. 16-26.
282. Pan, Y.R., et al., Analysis of Epstein-Barr virus gene expression upon phorbol ester and hydroxyurea treatment by real-time quantitative PCR. *Archives of Virology*, 2005. 150(4): p. 755-770.
283. Paschos, K. and M.J. Allday, Epigenetic reprogramming of host genes in viral and microbial pathogenesis. *Trends in Microbiology*, 2010. 18(10): p. 439-447.
284. Pasquale, E.B., Eph-ephrin bidirectional signaling in physiology and disease. *Cell*, 2008. 133(1): p. 38-52.
285. Paulson, E.J. and S.H. Speck, Differential methylation of Epstein-Barr virus latency promoters facilitates viral persistence in healthy seropositive individuals. *Journal of Virology*, 1999. 73(12): p. 9959-9968.
286. Pender, M.P. and J.M. Greer, Immunology of multiple sclerosis. *Current Allergy and Asthma Reports*, 2007. 7(4): p. 285-292.
287. Pender, M.P. and N.P. Wolfe, Prevention of autoimmune attack and disease progression in multiple sclerosis: current therapies and future prospects. *Intern Med J*, 2002. 32(11): p. 554-63.
288. Pender, M.P., The Essential Role of Epstein-Barr Virus in the Pathogenesis of Multiple Sclerosis. *Neuroscientist*, 2011. 17(4): p. 351-367.
289. Peral, B., et al., Comment: the methionine 196 arginine polymorphism in exon 6 of the TNF receptor 2 gene (TNFRSF1B) is associated with the polycystic ovary syndrome and hyperandrogenism. *J Clin Endocrinol Metab*, 2002. 87(8): p. 3977-83.
290. Pereira, J.P., L.M. Kelly, and J.G. Cyster, Finding the right niche: B-cell migration in the early phases of T-dependent antibody responses. *International Immunology*, 2010. 22(6): p. 413-419.
291. Peterfy, M., et al., Mutations in LMF1 cause combined lipase deficiency and severe hypertriglyceridemia. *Nature Genetics*, 2007. 39(12): p. 1483-1487.

292. Pfaffl, M.W., et al., Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. *Biotechnol Lett*, 2004. 26(6): p. 509-15.
293. Phillips, J.E. and V.G. Corces, CTCF: master weaver of the genome. *Cell*, 2009. 137(7): p. 1194-211.
294. Pickrell, J.K., et al., Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 2010. 464(7289): p. 768-772.
295. Piriou, E., et al., Early Age at Time of Primary Epstein-Barr Virus Infection Results in Poorly Controlled Viral Infection in Infants From Western Kenya: Clues to the Etiology of Endemic Burkitt Lymphoma. *Journal of Infectious Diseases*, 2012. 205(6): p. 906-913.
296. Preuschhof, C., et al., KIBRA and CLSTN2 polymorphisms exert interactive effects on human episodic memory. *Neuropsychologia*, 2010. 48(2): p. 402-8.
297. Price, A.M., et al., Analysis of Epstein-Barr Virus-Regulated Host Gene Expression Changes through Primary B-Cell Outgrowth Reveals Delayed Kinetics of Latent Membrane Protein 1-Mediated NF- $\kappa$ B Activation. *Journal of Virology*, 2012. 86(20): p. 11096-11106.
298. Pumplin, N. and O. Voinnet, RNA silencing suppression by plant pathogens: defence, counter-defence and counter-counter-defence. *Nat Rev Microbiol*, 2013. 11(11): p. 745-60.
299. Purcell, S., et al., PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 2007. 81(3): p. 559-75.
300. Radonic, A., et al., Guideline to reference gene selection for quantitative real-time PCR. *Biochem Biophys Res Commun*, 2004. 313(4): p. 856-62.
301. Ramagopalan, S.V., et al., Multiple sclerosis: risk factors, prodromes, and potential causal pathways. *Lancet Neurology*, 2010. 9(7): p. 727-739.
302. Ramdas, W.D., et al., A genome-wide association study of optic disc parameters. *PLoS Genet*, 2010. 6(6): p. e1000978.
303. Randazzo, P.A., et al., The Arf GTPase-activating protein ASAP1 regulates the actin cytoskeleton. *Proc Natl Acad Sci U S A*, 2000. 97(8): p. 4011-6.
304. Rao, A.V. and B.D. Smith, Are Results of Targeted Gene Sequencing Ready to Be Used for Clinical Decision Making for Patients with Acute Myelogenous Leukemia? *Current Hematologic Malignancy Reports*, 2013. 8(2): p. 149-155.
305. Rasti, N., M. Wahlgren, and Q.J. Chen, Molecular aspects of malaria pathogenesis. *Fems Immunology and Medical Microbiology*, 2004. 41(1): p. 9-26.
306. Repapi, E., et al., Genome-wide association study identifies five loci associated with lung function. *Nature Genetics*, 2010. 42(1): p. 36-44.
307. Repic, A.M., et al., Augmented Latent Membrane Protein 1 Expression from Epstein-Barr Virus Episomes with Minimal Terminal Repeats. *Journal of Virology*, 2010. 84(5): p. 2236-2244.
308. Resnick, M.A. and A. Inga, Functional mutants of the sequence-specific transcription factor p53 and implications for master genes of diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 2003. 100(17): p. 9934-9939.
309. Rezaei, N., et al., X-linked lymphoproliferative syndrome: a genetic condition typified by the triad of infection, immunodeficiency and lymphoma. *Br J Haematol*, 2011. 152(1): p. 13-30.
310. Rezk, S.A. and L.M. Weiss, Epstein-Barr virus-associated lymphoproliferative disorders. *Hum Pathol*, 2007. 38(9): p. 1293-304.

311. Rivat, C., et al., SAP gene transfer restores cellular and humoral immune function in a murine model of X-linked lymphoproliferative disease. *Blood*, 2013. 121(7): p. 1073-1076.
312. Robertson, E.S., Epstein-Barr virus : latency and transformation. 2010, Wymondham, Norfolk, UK: Caister Academic Press. viii, 200 p.
313. Roehl, A.C., et al., Extended runs of homozygosity at 17q11.2: an association with type-2 NF1 deletions? *Hum Mutat*, 2010. 31(3): p. 325-34.
314. Roland, M.E. and P.G. Stock, Review of solid-organ transplantation in HIV-infected patients. *Transplantation*, 2003. 75(4): p. 425-9.
315. Romeu, A. and L. Arola, Classical dynamin DNMI and DNMI3 genes attain maximum expression in the normal human central nervous system. *BMC Res Notes*, 2014. 7: p. 188.
316. Rooney, C., et al., Influence of Burkitt's lymphoma and primary B cells on latent gene expression by the nonimmortalizing P3J-HR-1 strain of Epstein-Barr virus. *Journal of Virology*, 1989. 63(4): p. 1531-9.
317. Rose, E.J., et al., Neural effects of the CSMD1 genome-wide associated schizophrenia risk variant rs10503253. *Am J Med Genet B Neuropsychiatr Genet*, 2013. 162B(6): p. 530-7.
318. Rosen, A., et al., Lymphoblastoid cell line with B1 cell characteristics established from a chronic lymphocytic leukemia clone by in vitro EBV infection. *Oncoimmunology*, 2012. 1(1): p. 18-27.
319. Rothenberg, S.M., et al., A Genome-Wide Screen for Microdeletions Reveals Disruption of Polarity Complex Genes in Diverse Human Cancers. *Cancer Res*, 2010. 70(6): p. 2158-2164.
320. Rubicz, R., et al., A genome-wide integrative genomic study localizes genetic factors influencing antibodies against Epstein-Barr virus nuclear antigen 1 (EBNA-1). *PLoS Genet*, 2013. 9(1): p. e1003147.
321. Rubio, J.P., et al., Replication of KIAA0350, IL2RA, RPL5 and CD58 as multiple sclerosis susceptibility genes in Australians. *Genes and Immunity*, 2008. 9(7): p. 624-630.
322. Ruf, S. and H.J. Wagner, Determining EBV load: current best practice and future requirements. *Expert Review of Clinical Immunology*, 2013. 9(2): p. 139-151.
323. Ruf, S., et al., EBV Load in Whole Blood Correlates With LMP2 Gene Expression After Pediatric Heart Transplantation or Allogeneic Hematopoietic Stem Cell Transplantation. *Transplantation*, 2014.
324. Ruibal-Ares, B.H., et al., HIV-1 infection and chemokine receptor modulation. *Current Hiv Research*, 2004. 2(1): p. 39-50.
325. S. Ocheni, D.B.O., A.A. Oyekunle, O.G. Ibegbulam, N. Kröger, U. Bacher and A.R. Zander, EBV-Associated Malignancies. *The Open Infectious Diseases Journal*, 2010(Special Issue): p. 11.
326. Said-Conti, V., et al., Successful treatment of central nervous system PTLD with rituximab and cranial radiotherapy. *Pediatr Nephrol*, 2013. 28(10): p. 2053-6.
327. Salzer, E., et al., Combined immunodeficiency with life-threatening EBV-associated lymphoproliferative disorder in patients lacking functional CD27. *Haematologica*, 2013. 98(3): p. 473-478.
328. Salzer, U., et al., Mutations in TNFRSF13B encoding TACI are associated with common variable immunodeficiency in humans. *Nature Genetics*, 2005. 37(8): p. 820-8.
329. Šantak, M., Identification and Functional Analysis of Epstein-Barr Nuclear Antigen 2 (EBNA2) Target Genes. 2004, Ludwig Maximilians Universität München.

330. Satoh, J. and H. Tabunoki, Molecular network of chromatin immunoprecipitation followed by deep sequencing-based vitamin D receptor target genes. *Multiple Sclerosis Journal*, 2013. 19(8): p. 1035-1045.
331. Sawcer, S., et al., Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 2011. 476(7359): p. 214-219.
332. Schlee, M., et al., Identification of Epstein-Barr virus (EBV) nuclear antigen 2 (EBNA2) target proteins by proteome analysis: Activation of EBNA2 in conditionally immortalized B cells reflects early events after infection of primary B cells by EBV. *Journal of Virology*, 2004. 78(8): p. 3941-3952.
333. Schmid, J.P., et al., Clinical similarities and differences of patients with X-linked lymphoproliferative syndrome type 1 (XLP-1/SAP deficiency) versus type 2 (XLP-2/XIAP deficiency). *Blood*, 2011. 117(5): p. 1522-1529.
334. Schneider, N., Association of cytokine gene polymorphisms with posttransplant lymphoproliferative disorder in Epstein-Barr positive transplant recipients, in *Department of Medicine - Charité - University Medicine Berlin*. 2013, University Medicine Berlin, Berlin.
335. Seidman, J.G. and C. Seidman, Transcription factor haploinsufficiency: when half a loaf is not enough. *J Clin Invest*, 2002. 109(4): p. 451-5.
336. Shabalin, A.A., Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 2012. 28(10): p. 1353-8.
337. Shaffer, A.L., et al., IRF4: Immunity. Malignancy! Therapy? *Clinical Cancer Research*, 2009. 15(9): p. 2954-2961.
338. Shah, K.M. and L.S. Young, Epstein-Barr virus and carcinogenesis: beyond Burkitt's lymphoma. *Clin Microbiol Infect*, 2009. 15(11): p. 982-8.
339. Shair, K.H., C.I. Schnegg, and N. Raab-Traub, Epstein-Barr virus latent membrane protein-1 effects on junctional plakoglobin and induction of a cadherin switch. *Cancer Res*, 2009. 69(14): p. 5734-42.
340. Shimoyama, Y., et al., Senile Epstein-Barr virus-associated B-cell lymphoproliferative disorders: a mini review. *J Clin Exp Hematop*, 2006. 46(1): p. 1-4.
341. Shiozawa, E., et al., Senile EBV-associated B-cell lymphoproliferative disorder of indolent clinical phenotype with recurrence as aggressive lymphoma. *Pathol Int*, 2007. 57(10): p. 688-93.
342. Shroff, R. and L. Rees, The post-transplant lymphoproliferative disorder - a literature review. *Pediatric Nephrology*, 2004. 19(4): p. 369-377.
343. Shull, A.Y., et al., Somatic mutations, allele loss, and DNA methylation of the Cub and Sushi Multiple Domains 1 (CSMD1) gene reveals association with early age of diagnosis in colorectal cancer patients. *PLoS One*, 2013. 8(3): p. e58731.
344. Sides, M.D., et al., The Epstein-Barr virus latent membrane protein 1 and transforming growth factor--beta1 synergistically induce epithelial--mesenchymal transition in lung epithelial cells. *Am J Respir Cell Mol Biol*, 2011. 44(6): p. 852-62.
345. Simon, K., et al., Variation in the Epstein-Barr virus receptor, CR2, and risk of multiple sclerosis. *Mult Scler*, 2007. 13(7): p. 947-8.
346. Sims, K., A. Saha, and E.S. Robertson, Regulation of Cellular Processes by the Epstein-Barr Virus Nuclear Antigen 3 Family of Proteins. *Epstein-Barr Virus: Latency and Transformation*, 2010: p. 81-100.
347. Sivachandran, N., J.Y. Cao, and L. Frappier, Epstein-Barr virus nuclear antigen 1 Hijacks the host kinase CK2 to disrupt PML nuclear bodies. *Journal of Virology*, 2010. 84(21): p. 11113-23.

348. Skibola, C.F., J.D. Curry, and A. Nieters, Genetic susceptibility to lymphoma. *Haematologica-the Hematology Journal*, 2007. 92(7): p. 960-969.
349. Slyker, J.A., et al., Clinical and Virologic Manifestations of Primary Epstein-Barr Virus (EBV) Infection in Kenyan Infants Born to HIV-Infected Women. *Journal of Infectious Diseases*, 2013. 207(12): p. 1798-1806.
350. Smets, F., et al., Ratio between Epstein-Barr viral load and anti-Epstein-Barr virus specific T-cell response as a predictive marker of posttransplant lymphoproliferative disease. *Transplantation*, 2002. 73(10): p. 1603-1610.
351. Smets, F., et al., Ratio between Epstein-Barr viral load and anti-Epstein-Barr virus specific T-cell response as a predictive marker of posttransplant lymphoproliferative disease. *Transplantation*, 2002. 73(10): p. 1603-1610.
352. Smith, C. and R. Khanna, Generation of Cytotoxic T Lymphocytes for Immunotherapy of EBV-Associated Malignancies. *Immunotherapy of Cancer: Methods and Protocols*, 2010. 651: p. 49-59.
353. Song, G.G., et al., Genome-wide pathway analysis of a genome-wide association study on multiple sclerosis. *Molecular Biology Reports*, 2013. 40(3): p. 2557-2564.
354. Soranzo, N., et al., A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genetics*, 2009. 41(11): p. 1182-90.
355. Spender, L.C. and G.J. Inman, Inhibition of germinal centre apoptotic programmes by epstein-barr virus. *Adv Hematol*, 2011. 2011: p. 829525.
356. Stahl, E.A., et al., Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics*, 2010. 42(6): p. 508-U56.
357. Steen, V.M., et al., Neuropsychological Deficits in Mice Depleted of the Schizophrenia Susceptibility Gene CSMD1. *PLoS One*, 2013. 8(11).
358. Stein, M.F., et al., Multiple Interferon Regulatory Factor and NF- $\kappa$ B Sites Cooperate in Mediating Cell-Type- and Maturation-Specific Activation of the Human CD83 Promoter in Dendritic Cells. *Mol Cell Biol*, 2013. 33(7): p. 1331-1344.
359. Stewart WCL, Cerise J. Increasing the power of association studies with affected families, unrelated cases and controls. *Frontiers in Genetics*2013;4:200. doi:10.3389/fgene.2013.00200.
360. Storey, Gene-expression variation within and among human populations (vol 80, pg 502, 2007). *American Journal of Human Genetics*, 2007. 80(6): p. 1194-1194.
361. Stranger, B.E., et al., Patterns of Cis Regulatory Variation in Diverse Human Populations. *Plos Genetics*, 2012. 8(4): p. 272-284.
362. Straus, S.E., et al., Epstein-Barr-Virus Infections - Biology, Pathogenesis, and Management. *Annals of Internal Medicine*, 1993. 118(1): p. 45-58.
363. Sugden, B., M. Phelps, and J. Domoradzki, Epstein-Barr Virus-DNA Is Amplified in Transformed Lymphocytes. *Journal of Virology*, 1979. 31(3): p. 590-595.
364. Sun, D., et al., Aberrant methylation of CDH13 gene in nasopharyngeal carcinoma could serve as a potential diagnostic biomarker. *Oral Oncology*, 2007. 43(1): p. 82-87.
365. Suzuki, R., et al., Extranodal NK/T-cell lymphoma: diagnosis and treatment cues. *Hematol Oncol*, 2008. 26(2): p. 66-72.
366. Suzuki, T., et al., Kaposi's sarcoma-associated herpesvirus-encoded LANA positively affects on ubiquitylation of p53. *Biochemical and Biophysical Research Communications*, 2010. 403(2): p. 194-197.
367. Swaminathan, S., The Role of Non-coding RNAs in EBV-induced Cell Growth and Transformation. *Epstein-Barr Virus: Latency and Transformation*, 2010: p. 155-166.
368. Swartz, M.E., et al., EphA4/ephrin-A5 interactions in muscle precursor cell migration in the avian forelimb. *Development*, 2001. 128(23): p. 4669-4680.

369. Takacs, M., et al., The importance of epigenetic alterations in the development of epstein-barr virus-related lymphomas. *Mediterr J Hematol Infect Dis*, 2009. 1(2): p. e2009012.
370. Takahashi, T., et al., Cutting edge: analysis of human V alpha 24+CD8+ NK T cells activated by alpha-galactosylceramide-pulsed monocyte-derived dendritic cells. *J Immunol*, 2002. 168(7): p. 3140-4.
371. Takahashi, T., et al., Immunologic self-tolerance maintained by CD25+CD4+ naturally anergic and suppressive T cells: induction of autoimmune disease by breaking their anergic/suppressive state. *Int Immunol*, 1998. 10(12): p. 1969-80.
372. Takeuchi, T., et al., Loss of T-cadherin (CDH13, H-cadherin) expression in cutaneous squamous cell carcinoma. *Laboratory Investigation*, 2002. 82(8): p. 1023-1029.
373. Tao, Q. and A.T. Chan, Nasopharyngeal carcinoma: molecular pathogenesis and therapeutic developments. *Expert Rev Mol Med*, 2007. 9(12): p. 1-24.
374. Tattevin, P., et al., Increasing incidence of severe Epstein-Barr virus-related infectious mononucleosis: surveillance study. *J Clin Microbiol*, 2006. 44(5): p. 1873-4.
375. Tempera, I., et al., CTCF Prevents the Epigenetic Drift of EBV Latency Promoter Qp. *Plos Pathogens*, 2010. 6(8).
376. Thacker, E.L., F. Mirzaei, and A. Ascherio, Infectious mononucleosis and risk for multiple sclerosis: a meta-analysis. *Ann Neurol*, 2006. 59(3): p. 499-503.
377. Thorley-Lawson, D.A. and M.J. Allday, The curious case of the tumour virus: 50 years of Burkitt's lymphoma. *Nature Reviews Microbiology*, 2008. 6(12): p. 913-24.
378. Thorley-Lawson, D.A., Epstein-Barr virus: exploiting the immune system. *Nature Reviews Immunology*, 2001. 1(1): p. 75-82.
379. Tierney, R.J., et al., Epstein-Barr-Virus Latency in Blood Mononuclear-Cells - Analysis of Viral Gene-Transcription during Primary Infection and in the Carrier State. *Journal of Virology*, 1994. 68(11): p. 7374-7385.
380. Titti, F., et al., Infection of simian B lymphoblastoid cells with simian immunodeficiency virus is associated with upregulation of CD23 and CD40 cell surface markers. *J Med Virol*, 2002. 68(1): p. 129-140.
381. Toyooka, S., et al., Aberrant methylation of the CDH13 (H-cadherin) promoter region in colorectal cancers and adenomas. *Cancer Research*, 2002. 62(12): p. 3382-3386.
382. Tsai, K., et al., EBV Tegument Protein BNRF1 Disrupts DAXX-ATRAX to Activate Viral Early Gene Transcription. *Plos Pathogens*, 2011. 7(11).
383. Tsao, S.W., et al., The biology of EBV infection in human epithelial cells. *Seminars in Cancer Biology*, 2012. 22(2): p. 137-143.
384. Tsuchiya, N., et al., Analysis of the association of HLA-DRB1, TNFalpha promoter and TNFR2 (TNFRSF1B) polymorphisms with SLE using transmission disequilibrium test. *Genes Immun*, 2001. 2(6): p. 317-22.
385. Turner, C.E., K.A. West, and M.C. Brown, Paxillin-ARF GAP signaling and the cytoskeleton. *Curr Opin Cell Biol*, 2001. 13(5): p. 593-9.
386. Tzartos, J.S., et al., Association of innate immune activation with latent Epstein-Barr virus in active MS lesions. *Neurology*, 2012. 78(1): p. 15-23.
387. Ulrich, H.D. and H. Walden, Ubiquitin signalling in DNA replication and repair. *Nat Rev Mol Cell Biol*, 2010. 11(7): p. 479-89.
388. Ura, K., et al., Enhanced RASGEF1A expression is involved in the growth and migration of intrahepatic cholangiocarcinoma. *Clinical Cancer Research*, 2006. 12(22): p. 6611-6616.
389. van Es, M.A., et al., Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nature Genetics*, 2008. 40(1): p. 29-31.

390. Van Hoecke, A., et al., EPHA4 is a disease modifier of amyotrophic lateral sclerosis in animal models and in humans. *Nature Medicine*, 2012. 18(9): p. 1418-+.
391. van Noort, J.M., et al., Mistaken self, a novel model that links microbial infections with myelin-directed autoimmunity in multiple sclerosis. *Journal of Neuroimmunology*, 2000. 105(1): p. 46-57.
392. Vanderlugt, C.L. and S.D. Miller, Epitope spreading in immunemediated diseases: Implications for immunotherapy. *Nature Reviews Immunology*, 2002. 2(2): p. 85-95.
393. Vandesompele, J., et al., Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*, 2002. 3(7).
394. Varade, J., et al., Replication study of 10 genes showing evidence for association with multiple sclerosis: validation of TMEM39A, IL12B and CLBL genes. *Multiple Sclerosis Journal*, 2012. 18(7): p. 959-965.
395. Veeramah, K.R., et al., De Novo Pathogenic SCN8A Mutation Identified by Whole-Genome Sequencing of a Family Quartet Affected by Infantile Epileptic Encephalopathy and SUDEP. *American Journal of Human Genetics*, 2012. 90(3): p. 502-510.
396. Veronese, M.L., et al., Detection of myc translocations in lymphoma cells by fluorescence in situ hybridization with yeast artificial chromosomes. *Blood*, 1995. 85(8): p. 2132-8.
397. Veyrieras, J.B., et al., High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*, 2008. 4(10): p. e1000214.
398. Vockerodt, M., et al., Suppression of the LMP2A target gene, EGR-1, protects Hodgkin's lymphoma cells from entry to the EBV lytic cycle. *J Pathol*, 2013. 230(4): p. 399-409.
399. Voight, B.F., et al., Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics*, 2010. 42(7): p. 579-89.
400. Wagner, M., et al., IL-12p70-dependent Th1 induction by human B cells requires combined activation with CD40 ligand and CpG DNA. *Journal of Immunology*, 2004. 172(2): p. 954-963.
401. Wakisaka, N., et al., Epstein-Barr virus latent membrane protein 1 induces synthesis of hypoxia-inducible factor 1 alpha. *Mol Cell Biol*, 2004. 24(12): p. 5223-34.
402. Wang, D., A. Rendon, and L. Wernisch, Transcription factor and chromatin features predict genes associated with eQTLs. *Nucleic Acids Res*, 2013. 41(3): p. 1450-63.
403. Wang, K.S., X.F. Liu, and N. Aragam, A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. *Schizophr Res*, 2010. 124(1-3): p. 192-9.
404. Wang, W.J., et al., Distinct Functional Effects for Dynamin 3 During Megakaryocytopoiesis. *Stem Cells and Development*, 2011. 20(12): p. 2139-2151.
405. Watson, J.D., *Molecular biology of the gene*. 6th ed. 2008, San Francisco Cold Spring Harbor, N.Y.: Pearson/Benjamin Cummings ;Cold Spring Harbor Laboratory Press. xxxii, 841 p.
406. Wei, Y.S., et al., Association of transforming growth factor-beta1 gene polymorphisms with genetic susceptibility to nasopharyngeal carcinoma. *Clin Chim Acta*, 2007. 380(1-2): p. 165-9.
407. Wei, Y.S., et al., Association of Variants in the Interleukin-27 and Interleukin-12 Gene With Nasopharyngeal Carcinoma. *Molecular Carcinogenesis*, 2009. 48(8): p. 751-757.

408. Weyn-Vanhentenryck, S.M., et al., HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Rep*, 2014. 6(6): p. 1139-52.
409. White, R.E., et al., Extensive Co-Operation between the Epstein-Barr Virus EBNA3 Proteins in the Manipulation of Host Gene Expression and Epigenetic Chromatin Modification. *Plos One*, 2010. 5(11).
410. Wildeman, M.A., et al., Short-term effect of different teaching methods on nasopharyngeal carcinoma for general practitioners in Jakarta, Indonesia. *PLoS One*, 2012. 7(3): p. e32756.
411. Willet, J.D., et al., Kidney transplantation: analysis of the expression and T cell-mediated activation of latent TGF-beta. *J Leukoc Biol*, 2013. 93(4): p. 471-8.
412. Williams, H., et al., Analysis of immune activation and clinical events in acute infectious mononucleosis. *Journal of Infectious Diseases*, 2004. 190(1): p. 63-71.
413. Wroblewski, J.M., et al., Cell surface phenotyping and cytokine production of Epstein-Barr Virus (EBV)-transformed lymphoblastoid cell lines (LCLs). *J Immunol Methods*, 2002. 264(1-2): p. 19-28.
414. Wu, D.Y., A. Krumm, and W.H. Schubach, Promoter-specific targeting of human SWI-SNF complex by Epstein-Barr virus nuclear protein 2. *Journal of Virology*, 2000. 74(19): p. 8893-8903.
415. Xia, K., et al., seeQTL: a searchable database for human eQTLs. *Bioinformatics*, 2012. 28(3): p. 451-452.
416. Yasui, Y., et al., Association of Epstein-Barr virus antibody titers with a human IL-10 promoter polymorphism in Japanese women. *J Autoimmune Dis*, 2008. 5: p. 2.
417. Yilmaz, V., S.P. Yentur, and G. Saruhan-Direskeneli, IL-12 and IL-10 polymorphisms and their effects on cytokine production. *Cytokine*, 2005. 30(4): p. 188-194.
418. Yoshizaki, T., et al., Current understanding and management of nasopharyngeal carcinoma. *Auris Nasus Larynx*, 2012. 39(2): p. 137-44.
419. Young, L.S. and A.B. Rickinson, Epstein-Barr virus: 40 years on. *Nature Reviews Cancer*, 2004. 4(10): p. 757-68.
420. Zamorano, J., et al., NF- $\kappa$ B activation plays an important role in the IL-4-induced protection from apoptosis. *International Immunology*, 2001. 13(12): p. 1479-1487.
421. Zandman-Goddard, G., et al., Exposure to Epstein-Barr Virus Infection Is Associated with Mild Systemic Lupus Erythematosus Disease. *Contemporary Challenges in Autoimmunity*, 2009. 1173: p. 658-663.
422. Zhang, W., et al., Mutation screening of the FKBP6 gene and its association study with spermatogenic impairment in idiopathic infertile men. *Reproduction*, 2007. 133(2): p. 511-516.
423. Zhang, Zhiwu, et al. "Mixed linear model approach adapted for genome-wide association studies." *Nature genetics* 42.4 (2010): 355-360.
424. Zhao, B., et al., Epstein-Barr virus nuclear antigen 3C regulated genes in lymphoblastoid cell lines. *Proc Natl Acad Sci U S A*, 2011. 108(1): p. 337-42.
425. Zhao, J., et al., Cadherin-12 contributes to tumorigenicity in colorectal cancer by promoting migration, invasion, adhesion and angiogenesis. *J Transl Med*, 2013. 11: p. 288.
426. Zhao, Y. and Y. Wang, 5T4 oncotrophoblast glycoprotein: janus molecule in life and a novel potential target against tumors. *Cell Mol Immunol*, 2007. 4(2): p. 99-104.
427. Zinkernagel, A.S., R.S. Johnson, and V. Nizet, Hypoxia inducible factor (HIF) function in innate immunity and infection. *J Mol Med (Berl)*, 2007. 85(12): p. 1339-46.

## Appendix

Table A1 – HapMap IDs for the 58 YRI LCLs from the hypoxia response study (Mohr 2010)

### 58 LCLs from Mohr (2010)

NA18502  
NA18505  
NA18508  
NA18501  
NA18517  
NA18523  
NA18522  
NA18871  
NA18853  
NA18855  
NA18856  
NA18861  
NA19137  
NA19171  
NA19210  
NA19206  
NA19207  
NA19160  
NA19130  
NA19193  
NA18504  
NA18859  
NA18870  
NA18912  
NA18913  
NA19093  
NA19102  
NA19101  
NA19138  
NA19201  
NA19200  
NA19172  
NA19203  
NA19159  
NA19099  
NA19116  
NA19128  
NA19131

NA19140  
 NA19141  
 NA19143  
 NA19153  
 NA19192  
 NA19222  
 NA18507  
 NA18858  
 NA18516  
 NA18852  
 NA19092  
 NA19204  
 NA19209  
 NA19098  
 NA19127  
 NA19144  
 NA19152  
 NA19223  
 NA19238  
 NA19239

**Figure A1 – WDR48 regional association plot.**

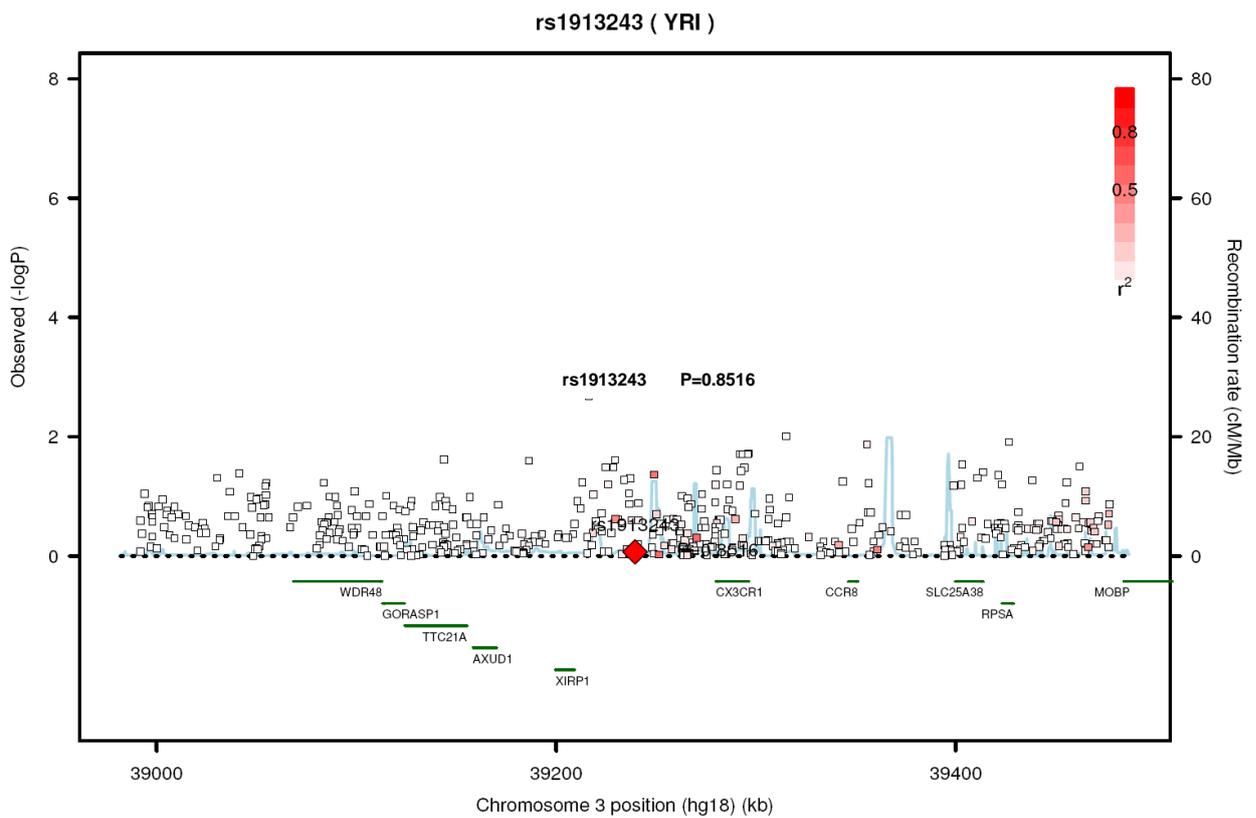


Figure A2 – CX3CR1 regional association plot.

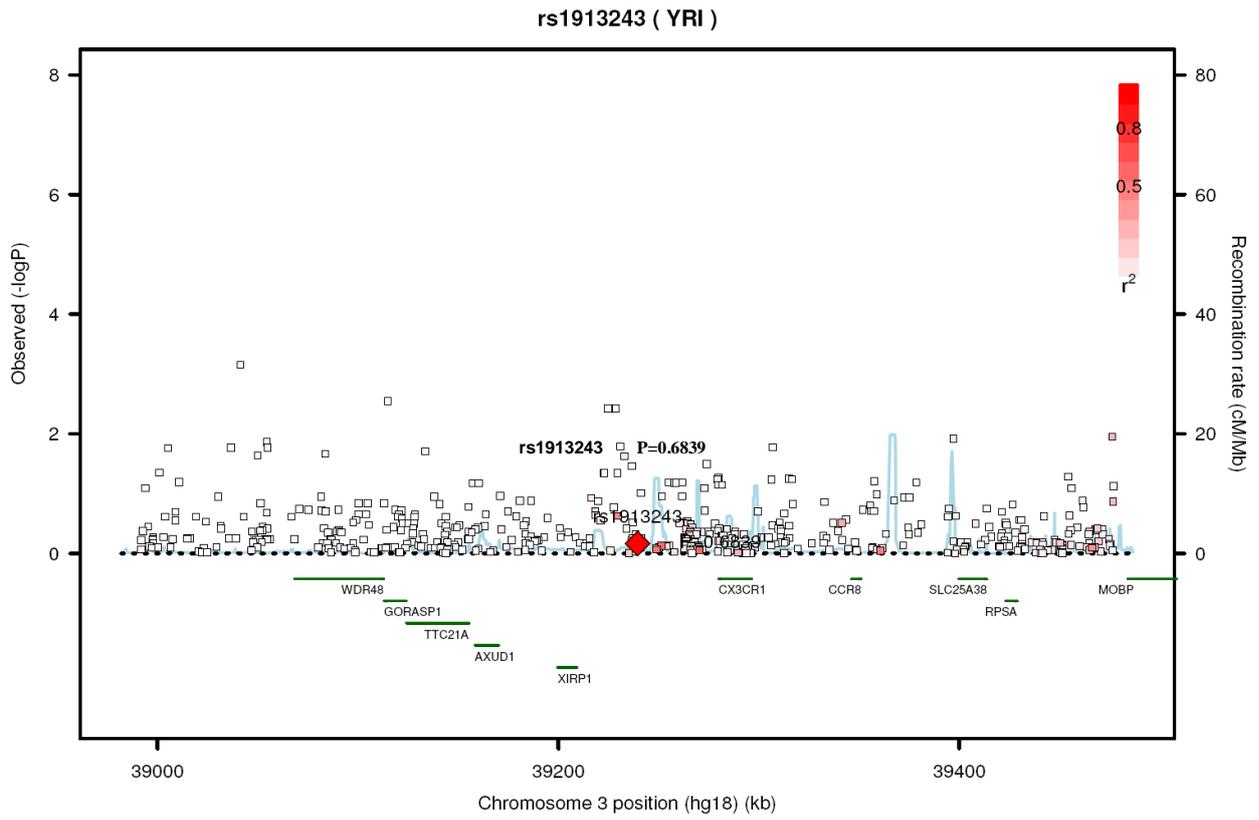
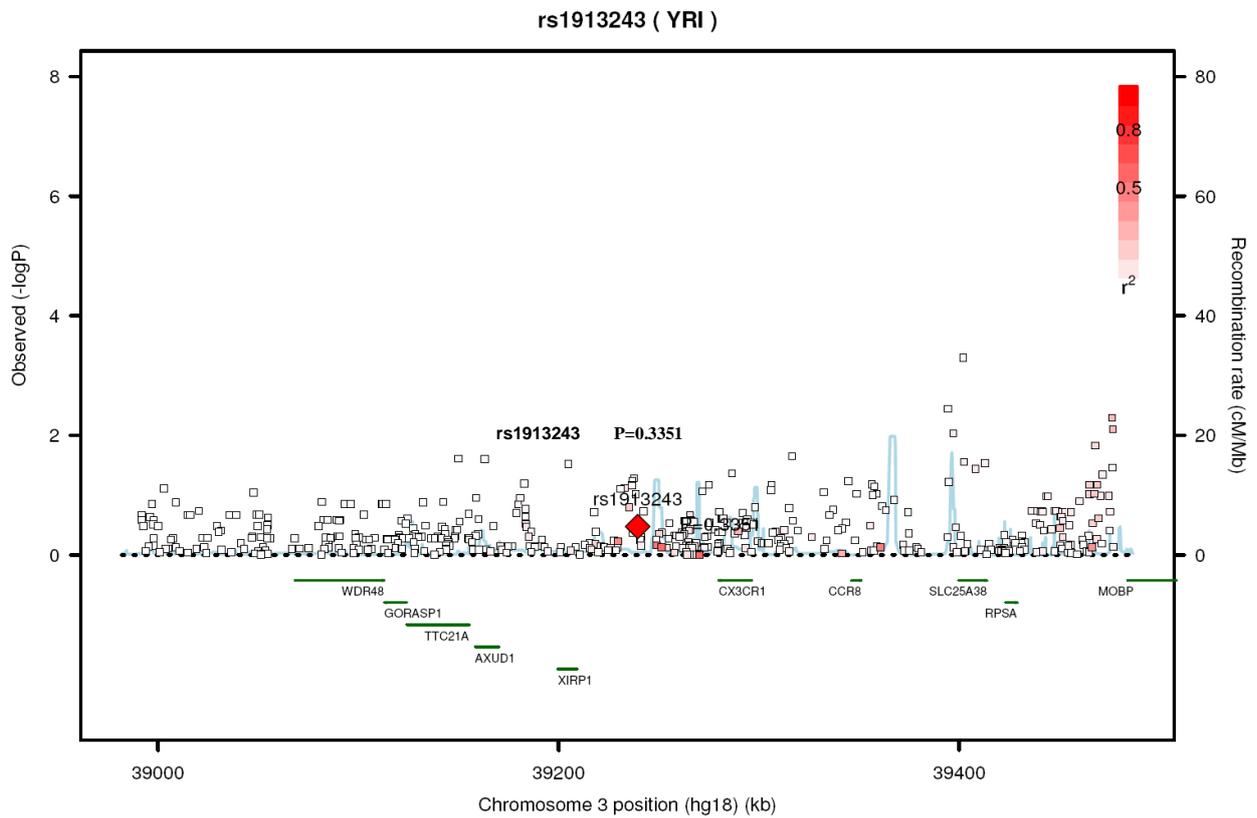
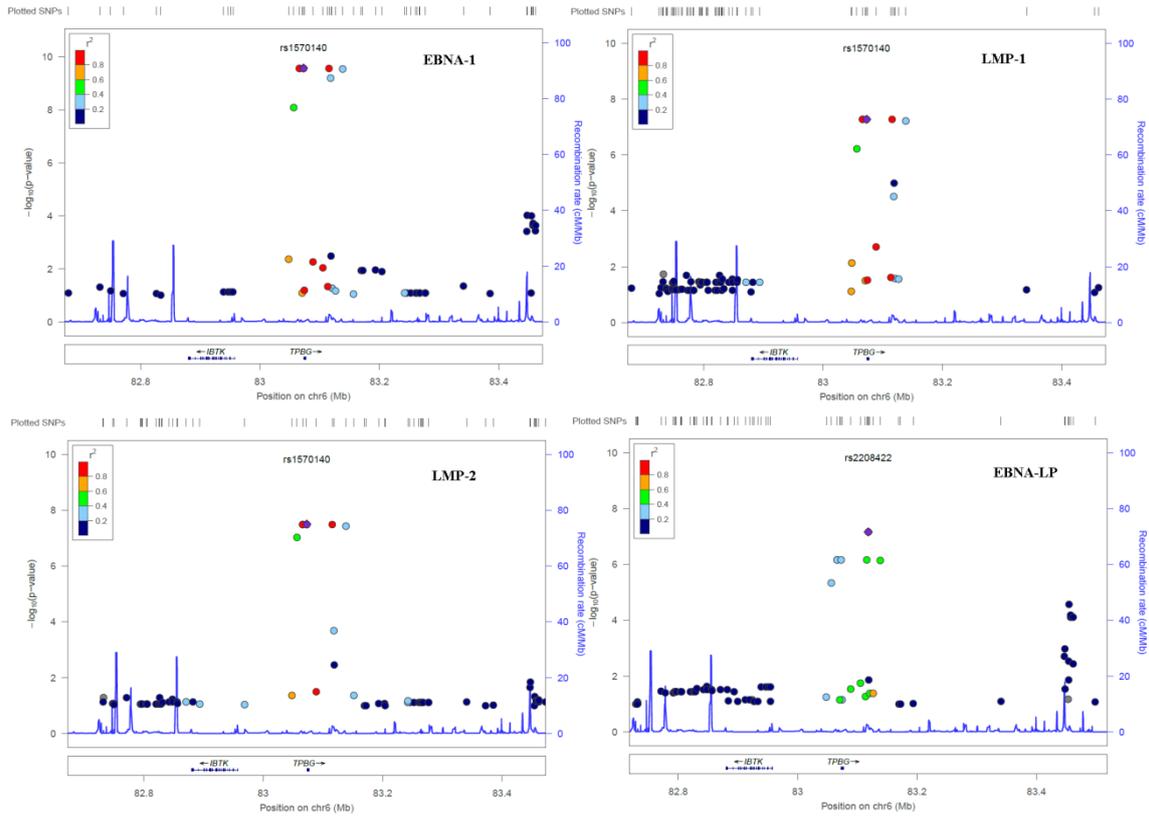


Figure A3 –XIRP1 regional association plot.



**Figure A4** – Regional association plots for 4 EBV latency transcripts (indicated within the plot’s field) and TPBG locus. Only associations with  $p < 1E-01$  are shown.



**Table A2** – Number of samples used for each t-test in the follow up experiment from chapter 6.

	Naive	Unstimulated	Stimulated	LCL week 2	LCL mature
CD83	27	39	37	23	23
CC / CT&TT	24/8	25/11	22/8	17/5	15/4
RASGEF1A	35	31	27	23	23
CC / CT&TT	21/9	15/11	15/10	12/8	13/7
IL-12B	32	32	32	23	23
CC / CT&TT	17/13	20/9	19/12	17/4	15/8
EBNA1	-	-	-	23	23
CC / CT&TT	-	-	-	12 / 5	13 / 4
EBNA3A	-	-	-	23	23
CC / CT&TT	-	-	-	15 / 8	15 / 8

Table A3 –MRCAs cohort samples provided for the EBV QTL project. Availability of global expression data as well as Illumina Sentrix HumanHap300 BeadChip (ILMN300K) and Illumina Sentrix Human-1 Genotyping BeadChip (ILMN100K) genotypes indicated.

Sample ID	Affy Expression data	100k genotypes	300k genotypes
8201.3	YES	YES	
8207.3	YES	YES	YES
8204.3	YES	YES	
8196.3	YES	YES	YES
8201.5	YES	YES	
8202.3	YES	YES	
8197.4	YES	YES	
8202.4	YES	YES	
8197.3	YES	YES	YES
8207.4	YES	YES	
8204.4	YES	YES	
8203.3	YES	YES	
8200.4	YES	YES	YES
8209.3	YES	YES	
8209.4	YES	YES	
8201.4	YES	YES	
8067.4	YES	YES	YES
8039.4	YES	YES	YES
8052.3	YES	YES	YES
8036.3	YES	YES	YES
8046.3	YES	YES	YES
8060.5	YES	YES	YES
8062.3	YES	YES	YES
8080.5	YES	YES	YES
8039.3	YES	YES	YES
8081.3	YES	YES	YES
8006.4	YES	YES	YES
8067.3	YES	YES	YES
8062.4	YES	YES	YES
8033.4	YES	YES	YES
8087.5	YES	YES	YES
8096.3	YES	YES	YES
8060.3	YES	YES	YES
8080.3	YES	YES	
8087.3	YES	YES	YES
8046.4	YES	YES	YES
8052.4	YES	YES	YES
8111.5	YES	YES	YES

8111.3	YES	YES	YES
8131.5	YES	YES	
8127.3	YES	YES	YES
8100.3	YES	YES	YES
8110.3	YES	YES	YES
8099.4	YES	YES	
8099.3	YES	YES	YES
8001.4	YES	YES	YES
8127.5	YES	YES	YES
8135.3	YES	YES	YES
8131.3	YES	YES	YES
8122.4	YES	YES	
8183.4	YES	YES	YES
8136.3	YES	YES	YES
8089.3	YES	YES	YES
8146.3	YES	YES	YES
8136.4	YES	YES	YES
8183.3	YES	YES	YES
8089.4	YES	YES	YES
8148.4	YES	YES	YES
8158.3	YES	YES	YES
8156.4	YES	YES	YES
8100.4	YES	YES	YES
8148.5	YES	YES	YES
8187.3	YES	YES	YES
8148.3	YES	YES	YES
8189.4	YES	YES	YES
8154.4	YES	YES	YES
8110.4	YES	YES	YES
8135.4	YES	YES	YES
8156.3	YES	YES	YES
8198.4	YES	YES	YES
8208.4	YES	YES	YES
8208.3	YES	YES	
8154.3	YES	YES	YES
8199.3	YES	YES	YES
8182.3	YES	YES	YES
8182.4	YES	YES	
8189.3	YES	YES	
8198.3	YES	YES	YES
8205.3	YES	YES	
8199.5	YES	YES	YES
8068.4	YES	YES	YES
8069.3	YES	YES	YES
8069.4	YES	YES	YES
8075.5	YES	YES	YES

8076.3	YES	YES	YES
8076.4	YES	YES	YES
8199.4	YES	YES	YES
8061.3	YES	YES	YES
8065.4	YES	YES	YES
8066.3	YES	YES	YES
8078.3	YES	YES	YES
8063.4	YES	YES	YES
8064.4	YES	YES	YES
8065.3	YES	YES	YES
8072.3	YES	YES	YES
8073.3	YES	YES	YES
8073.4	YES	YES	YES
8073.5	YES	YES	YES
8075.3	YES	YES	YES
8098.4	YES	YES	YES
8094.4	YES	YES	YES
8104.4	YES	YES	YES
8101.6	YES	YES	YES
8091.4	YES	YES	YES
8105.3	YES	YES	YES
8094.3	YES	YES	
8101.5	YES	YES	YES
8092.3	YES	YES	YES
8108.4	YES	YES	YES
8098.3	YES	YES	YES
8104.3	YES	YES	YES
8105.4	YES	YES	YES
8090.4	YES	YES	YES
8090.3	YES	YES	YES
8176.4	YES	YES	YES
8081.4	YES	YES	YES
8158.4	YES	YES	YES
8179.4	YES	YES	YES
8169.3	YES	YES	
8169.4	YES	YES	
8176.3	YES	YES	YES
8179.3	YES	YES	YES
8107.4	YES	YES	YES
8112.4	YES	YES	YES
8112.3	YES	YES	YES
8172.3	YES	YES	YES
8108.5	YES	YES	YES
8172.4	YES	YES	
8075.4	YES	YES	YES
8167.4	YES	YES	YES

8128.3	YES	YES	YES
8136.5	YES	YES	
8135.5	YES	YES	YES
8146.4	YES	YES	YES
8061.4	YES	YES	YES
8116.3	YES	YES	YES
8109.3	YES	YES	YES
8177.4	YES	YES	
8174.5	YES	YES	YES
8213.4	YES	YES	
8180.4	YES	YES	YES
8180.3	YES	YES	YES
8175.4	YES	YES	YES
8177.3	YES	YES	YES
8177.5	YES	YES	YES
8213.3	YES	YES	YES
8213.5	YES	YES	
8211.4	YES	YES	
8178.4	YES	YES	YES
8175.5	YES	YES	YES
8214.4	YES	YES	
8126.4	YES	YES	YES
8132.3	YES	YES	YES
8120.3	YES	YES	YES
8129.4	YES	YES	YES
8119.3	YES	YES	YES
8175.3	YES	YES	YES
8211.3	YES	YES	
8145.3	YES	YES	YES
8151.4	YES	YES	YES
8121.3	YES	YES	YES
8139.4	YES	YES	YES
8140.4	YES	YES	
8141.3	YES	YES	YES
8133.3	YES	YES	YES
8133.4	YES	YES	YES
8142.4	YES	YES	YES
8138.3	YES	YES	YES
8147.3	YES	YES	YES
8137.3	YES	YES	YES
8138.4	YES	YES	YES
8145.4	YES	YES	YES
8144.3	YES	YES	YES
8130.4	YES	YES	YES
8118.6	YES	YES	YES
8117.4	YES	YES	YES

8129.3	YES	YES	YES
8181.3	YES	YES	YES
8127.4	YES	YES	YES
8118.3	YES	YES	YES
8125.5	YES	YES	YES
8125.4	YES	YES	YES
8125.3	YES	YES	YES
8118.5	YES	YES	YES
8130.3	YES	YES	
8121.4	YES	YES	YES
8157.3	YES	YES	YES
8165.4	YES	YES	YES
8200.3	YES	YES	YES
8166.4	YES	YES	YES
8168.4	YES	YES	YES
8210.4	YES	YES	
8152.4	YES	YES	YES
8170.4	YES	YES	
8153.3	YES	YES	YES
8163.4	YES	YES	YES
8150.4	YES	YES	YES
8143.5	YES	YES	YES
8156.5	YES	YES	YES
8160.4	YES	YES	YES
8152.3	YES	YES	YES
8161.4	YES	YES	YES
8167.3	YES	YES	YES
8155.4	YES	YES	YES
8170.3	YES	YES	YES
8163.3	YES	YES	YES
8161.3	YES	YES	YES
8171.3	YES	YES	
8042.3	YES	YES	YES
8172.5	YES	YES	
8153.4	YES	YES	YES
8080.4	YES	YES	YES
8064.3	YES	YES	YES
8159.4	YES	YES	YES
8083.3	YES	YES	YES
8169.5	YES	YES	
8091.3	YES	YES	YES
8072.4	YES	YES	YES
8173.3	YES	YES	YES
8155.3	YES	YES	
8174.4	YES	YES	YES
8162.3	YES	YES	YES

8165.3	YES	YES	YES
8162.4	YES		
8163.5	YES	YES	YES
8157.5	YES	YES	YES
8020.3	YES	YES	YES
8020.4	YES	YES	YES
8055.3	YES	YES	YES
8023.4	YES	YES	YES
8005.4	YES	YES	YES
8015.3	YES	YES	YES
8059.3	YES	YES	YES
8044.3	YES	YES	YES
8027.4	YES	YES	
8025.4	YES	YES	YES
8030.4	YES	YES	YES
8201.6	YES	YES	
8017.4	YES	YES	YES
8009.3	YES	YES	YES
8008.3	YES	YES	YES
8028.4	YES	YES	YES
8040.4	YES	YES	YES
8047.3	YES	YES	YES
8045.3	YES	YES	YES
8049.3	YES	YES	YES
8045.4	YES	YES	YES
8054.4	YES	YES	YES
8060.4	YES	YES	YES
8034.3	YES	YES	YES
8190.3	YES		
8111.4	YES	YES	
8008.4	YES	YES	YES
8058.4	YES	YES	YES
8017.3	YES	YES	YES
8019.3	YES	YES	YES
8003.5	YES	YES	YES
8021.3	YES	YES	YES
8019.4	YES	YES	YES
8004.3	YES	YES	YES
8028.3	YES	YES	YES
8058.3	YES	YES	YES
8021.4	YES	YES	YES
8038.4	YES	YES	YES
8057.4	YES	YES	YES
8015.5	YES	YES	
8037.3	YES	YES	YES
8022.4	YES	YES	YES

8047.4	YES	YES	YES
8007.3	YES	YES	YES
8030.3	YES	YES	YES
8041.4	YES	YES	YES
8054.5	YES	YES	YES
8043.4	YES	YES	YES
8022.5	YES	YES	YES
8013.4	YES	YES	YES
8143.3	YES	YES	YES
8055.4	YES	YES	YES
8007.4	YES	YES	YES
8057.3	YES	YES	YES
8084.4	YES	YES	YES
8002.4	YES	YES	
8203.4	YES	YES	YES
8026.4	YES	YES	YES
8118.4	YES	YES	YES
8044.4	YES	YES	YES
8103.3	YES	YES	YES
8037.4	YES	YES	YES
8188.3	YES	YES	YES
8195.3	YES		
8034.4	YES	YES	
8194.4	YES	YES	YES
8026.3	YES	YES	YES
8049.4	YES	YES	YES
8141.4	YES	YES	YES
8043.3	YES	YES	YES
8186.4	YES	YES	YES
8191.3	YES	YES	YES
8192.3	YES	YES	YES
8193.4	YES	YES	YES
8210.3	YES	YES	
8192.4	YES	YES	
8150.3	YES	YES	YES
8054.3	YES	YES	YES
8012.4	YES	YES	YES
8166.3	YES	YES	YES
8042.4	YES	YES	YES
8168.3	YES	YES	YES
8013.3	YES	YES	YES
8010.3	YES	YES	YES
8014.3	YES	YES	YES
8196.5	YES	YES	YES
8079.3	YES	YES	YES
8165.5	YES	YES	YES

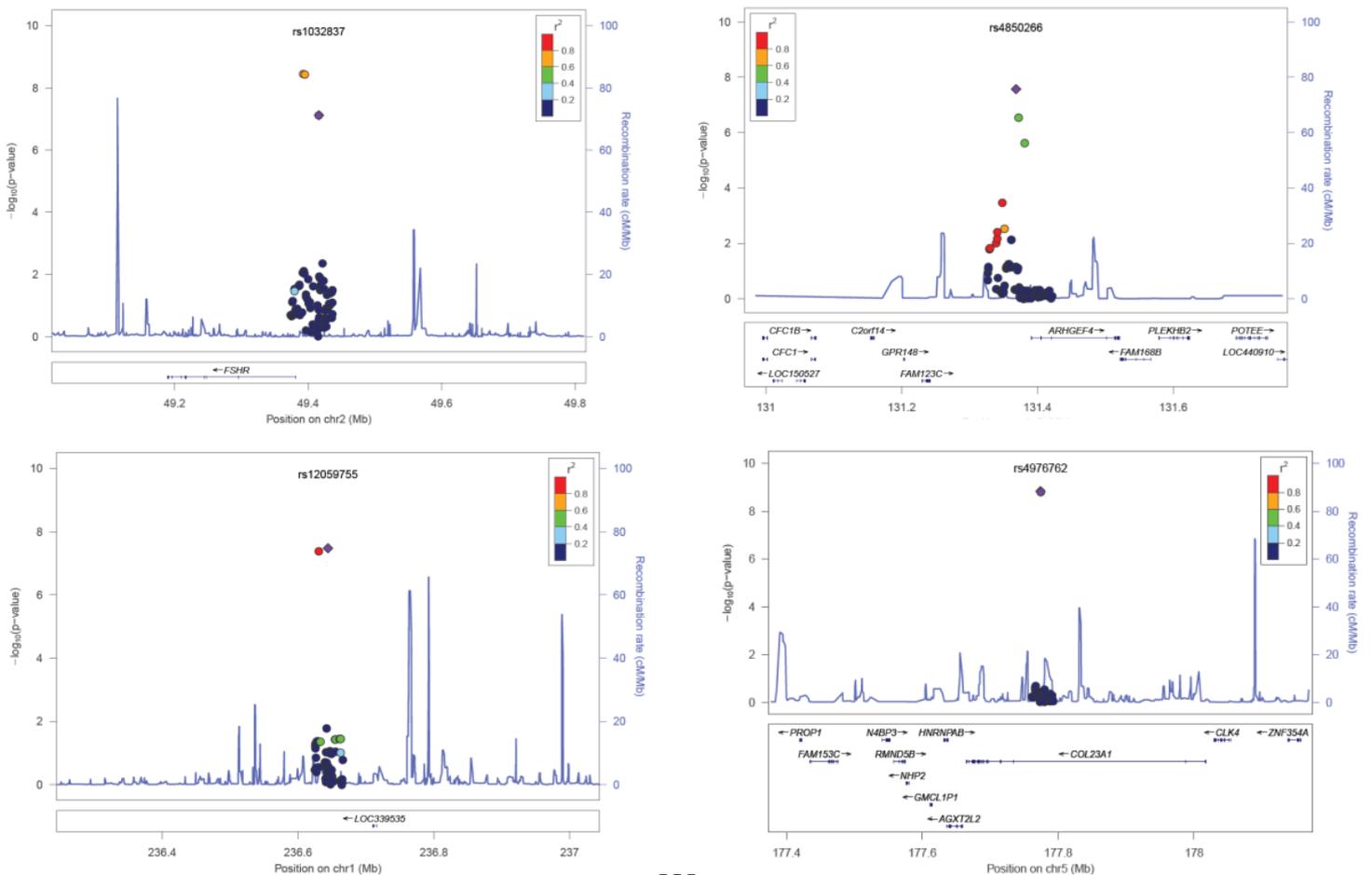
8174.3	YES	YES	YES
8184.5	YES	YES	YES
8151.3	YES	YES	YES
8113.4	YES	YES	YES
8190.5	YES	YES	
8191.4	YES	YES	YES
8188.4	YES	YES	YES
8184.4	YES	YES	YES
8196.4	YES	YES	YES
8063.3	YES	YES	YES
8048.3	YES	YES	YES
8041.3	YES	YES	YES
8101.4	YES	YES	YES
8051.3	YES	YES	YES
8185.4	YES		
8206.4	YES	YES	
8149.3	YES	YES	YES
8038.3	YES	YES	YES
8088.4	YES	YES	YES
8006.3	YES	YES	YES
8206.3	YES	YES	
8033.3	YES	YES	YES
8096.4	YES	YES	YES
8016.3	YES	YES	YES
8040.3	YES	YES	YES
8050.3	YES	YES	YES
8050.4	YES	YES	YES
8053.3	YES	YES	YES
8115.3	YES	YES	
8134.4	YES	YES	YES
8123.3	YES	YES	YES
8083.5	YES	YES	YES
8059.4	YES	YES	YES
8124.5	YES	YES	YES
8124.4	YES	YES	YES
8147.4	YES	YES	
8124.3	YES	YES	YES
8171.4	YES	YES	YES

**Table A4** – EBV latency eQTL candidates obtained using log-transformation of expression Ct-values. Table depicts results of a genome-wide association test for 10 latency transcripts in 277 LCLs derived from individuals of White British descent from the MRCA panel. Statistically significant associations (SNPs) are listed in the first column. In the second field “+” sign marks cases when a SNP has been also identified as an eQTLs for a human transcript by a global expression assay conducted on the same MRCA panel. “+” sign in the “PLOT” column indicates that more than 3 SNPs appeared to be associated with the expression and therefore a regional association plot could be constructed. “+” in the fourth column indicates MAF of >5%. The fifth and the sixth columns indicate the tested phenotype and its corresponding P-value, while the next field shows the gene in which the SNP is located in. The final field lists EBV-relevant eQTLs associated with a given SNP from SCAN database. Top candidates, marked in red, were further explored in a follow-up experiment. Grey background indicates close SNPs located within 100kb from each other.

SNP	eQTL?	PLOT	MAF	TRAIT	PVALUE	gene	eQTL / disease association
rs2444317				EBNA3B	5.88E-10		
rs12081575				EBNA3B	1.23E-09		
<b>rs4976762</b>		+		<b>EBNA3B</b>	<b>1.48E-09</b>	<b>COL23A1</b>	
<b>rs17159070</b>		+		<b>EBNA3B</b>	<b>1.95E-09</b>		
<b>rs13427781</b>		+	+	<b>LMP2</b>	<b>3.53E-09</b>		
<b>rs13428062</b>		+	+	<b>LMP2</b>	<b>3.58E-09</b>		
<b>rs7982202</b>		+		<b>EBNA3B</b>	<b>4.83E-09</b>		
rs12583649				EBNA3B	4.85E-09		
rs4868911				EBNA3B	4.90E-09		
rs13081968				EBNA3B	6.06E-09	BBX	
rs12533416				EBNA3B	6.09E-09	FLJ14712	
rs7196035				EBNA3B	6.99E-09	WVOX	B-cell lymphoma
rs16859293				EBNA3B	7.23E-09	TRIB2	MS-related
rs13301867				EBNA3B	7.56E-09	LOC644620	
<b>rs13331052</b>		+		<b>EBNA3B</b>	<b>1.00E-08</b>		
rs17158616			+	EBNA1	1.02E-08	RASGEF1A	
rs17158630			+	EBNA1	1.31E-08	RASGEF1A	eQTL for CD83 - interacts with EBV's LMP1
rs17066880				EBNA3B	1.51E-08	CSMD1	MS-associated in a GWAS, but not replicated
rs17339199			+	EBNA3A	1.88E-08	FKBP6	qQTL for IL12B - MS and lymphoma associated
<b>rs4850266</b>		+		<b>EBNA3C</b>	<b>2.68E-08</b>	<b>LOC730032</b>	
<b>rs12059755</b>		+		<b>EBNA3B</b>	<b>3.33E-08</b>		
rs316145				EBNA3B	3.41E-08	ICK	
rs10495408				EBNA3B	4.08E-08		
rs7748175				EBNA3B	5.14E-08		CNNM2 AGPAT7
rs13199042				EBNA3A	5.70E-08		
rs978136				EBNA3C	5.87E-08	LHFP	
rs17202166				EBNA3A	5.94E-08		
<b>rs6903311</b>		+		<b>LMP2</b>	<b>6.33E-08</b>	<b>MDGA1</b>	

rs13331864			EBNA3A	6.46E-08	
rs11867159	-	+	EBER1	5.33E-08	A2BP1/RBFOX1
rs12660137	+		EBER1	7.57E-08	eQTL for TNFRSF1B - MS SLE PTLD associated
<b>rs7196141</b>			<b>EBNA3A</b>	<b>7.09E-08</b>	
rs17487056			EBNA3B	7.10E-08	
<b>rs1032837</b>	+	+	<b>LMP2</b>	<b>7.56E-08</b>	
rs7913368			EBNA3B	7.98E-08	PTP4A3- activated by EBV
rs10457406			EBNA3B	8.11E-08	
rs7810849	+		EBNA3B	8.26E-08	
rs2717171			EBNA3B	8.43E-08	
<b>rs17865071</b>	+		<b>EBNA3B</b>	<b>8.44E-08</b>	eQTL for HJURP - upregulated by EBV
rs6107936			EBNA3B	8.79E-08	
<b>rs6077140</b>	+		<b>EBNA3B</b>	<b>9.08E-08</b>	
rs6519372			EBNA3B	9.16E-08	
rs17865486			EBNA3B	9.34E-08	

Figure A5 – Simplified regional plots constructed in LocusZoom for the chosen associations from table A4 (indicated in bold).



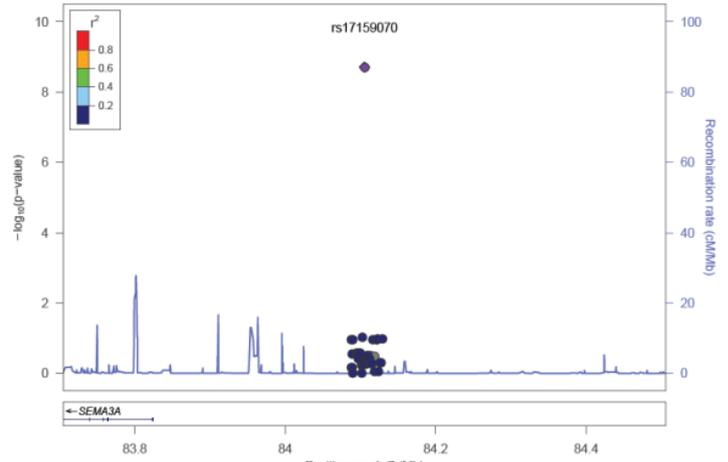
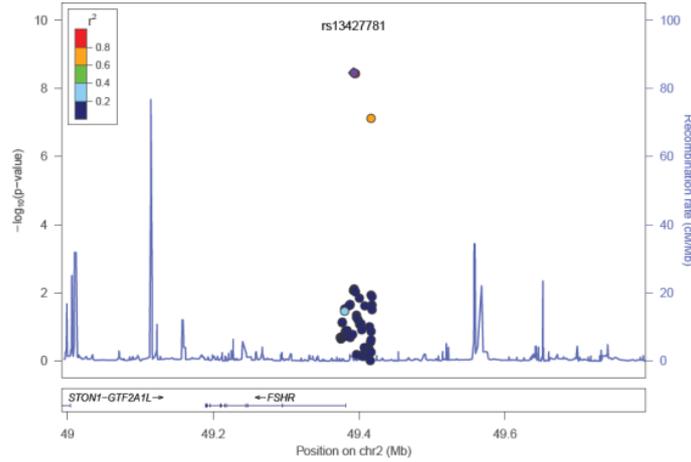
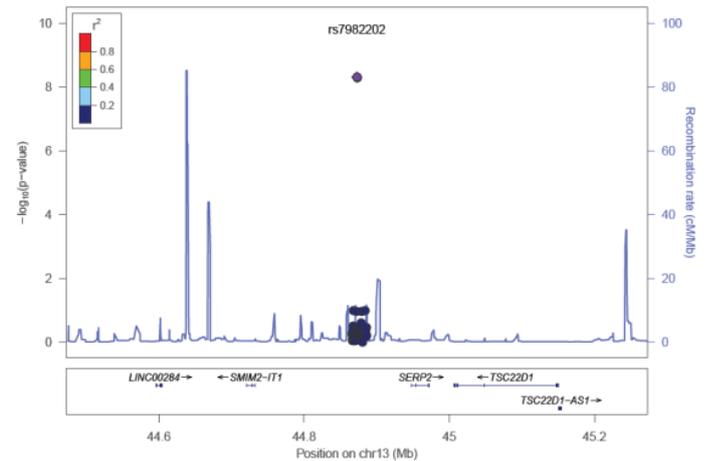
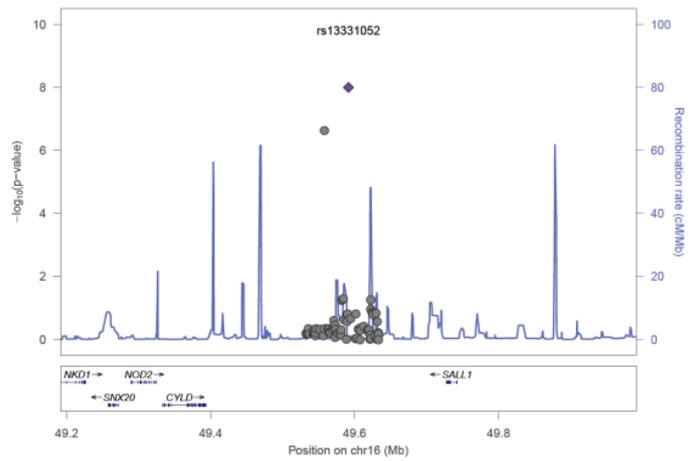
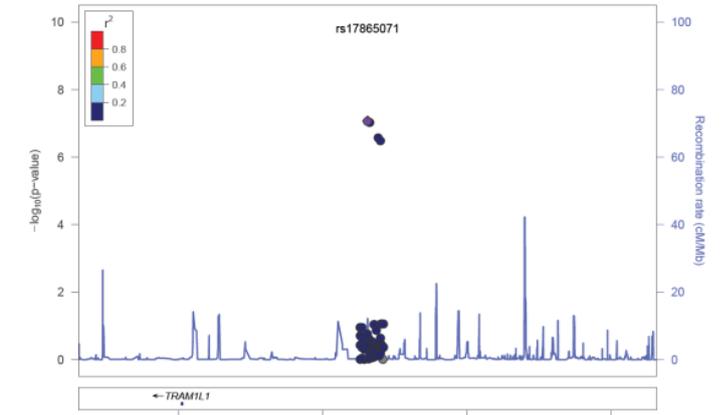
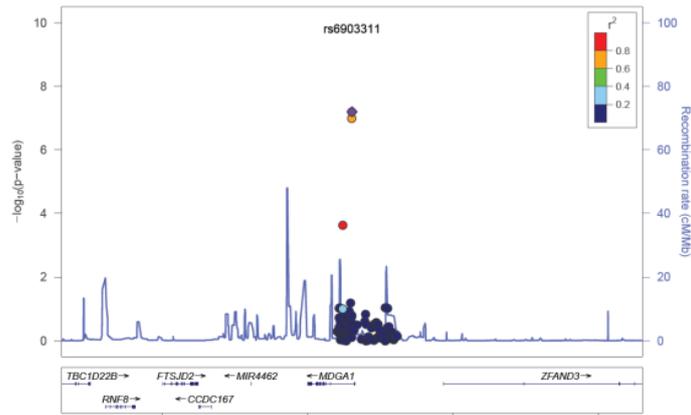
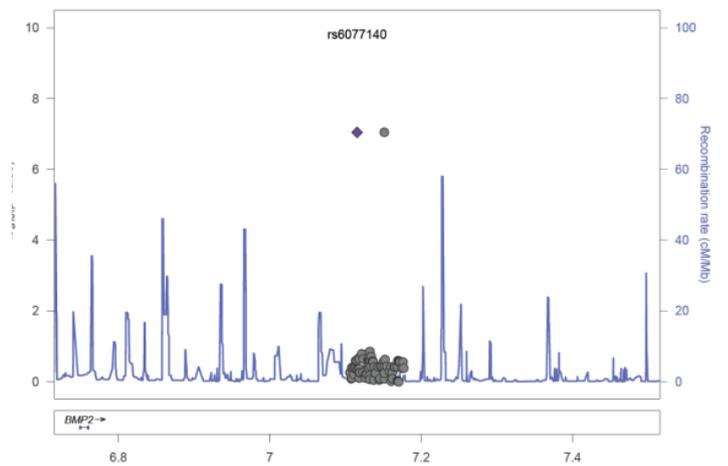
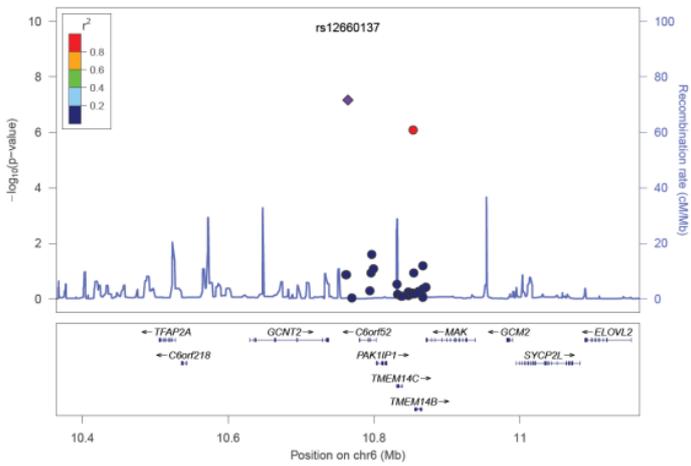


Figure A6 – EBV copy number associations in the 1000G Human Variation panel samples. The plot depicts associations for Chromosome 5 obtained from a test conducted on log-transformed phenotypes and is identical to Figure 75.

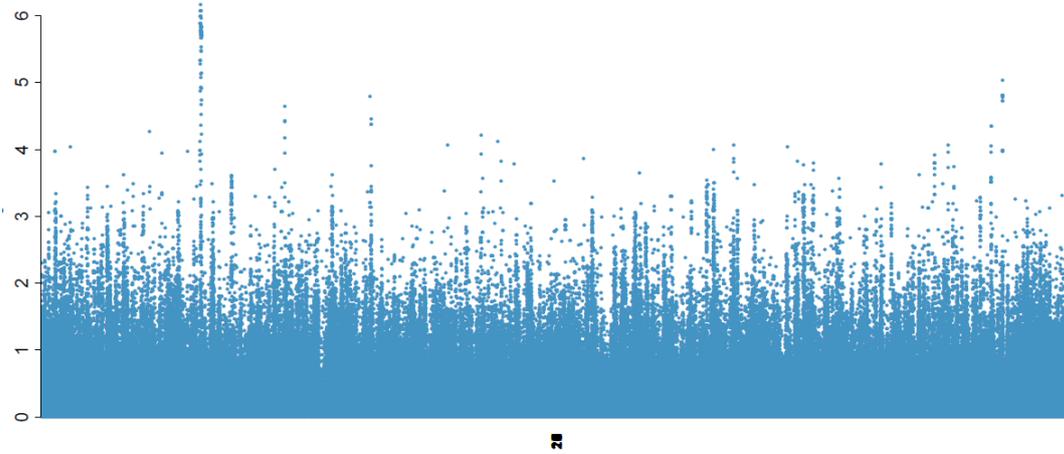
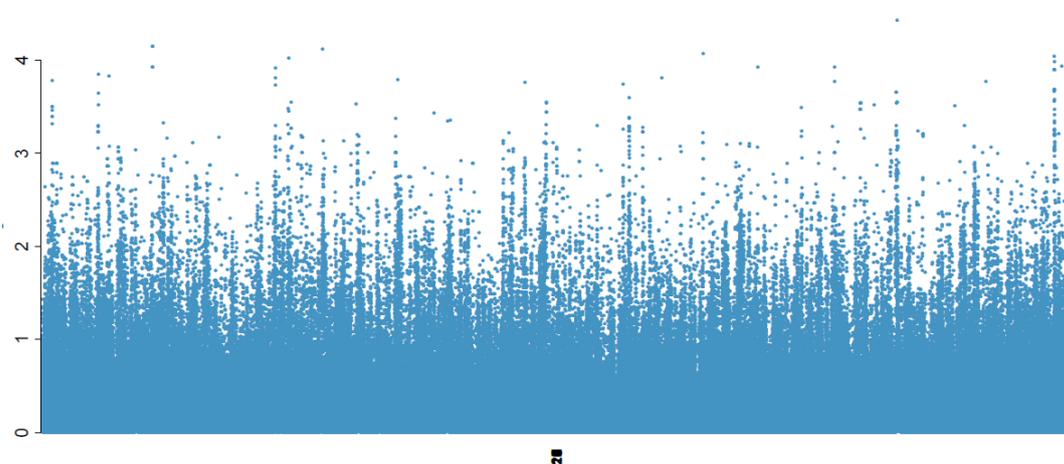


Figure A7 – EBV copy number association in the 1000G Human Variation panel samples. This plot depicts the same associations as Figure A6, however obtained using quantile-normalised phenotypes.



## MDS analysis

Run the following commands on the chosen ped/map/fam file to get MDS values for all the samples:

```
plink --file .... --read-genome  
...genome --cluster --mds-plot 4 --extract
```

Use the MDS file which is created as the output. Use the C1, C2, C3, and C4 columns to create the following graphs in R:

- C1 vs C2
- C1 vs C3
- C2 vs C3
- C3 vs C4

Identify outliers for removal from the other QC steps.