The Major Histocompatibility Complex Class I in the Pathogenesis of B-Cell Lymphomas

Karen Gomez

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

The Major Histocompatibility Complex Class I in the Pathogenesis of B-Cell Lymphomas

Karen Gomez

Immune evasion is an emerging hallmark of cancer. Dysregulation of the major histocompatibility complex class I (MHC-I) is a frequent mechanism of immune evasion utilized by tumor cells and is particularly relevant to the pathogenesis of B-cell lymphomas, including diffuse large B-cell lymphoma (DLBCL) and classical Hodgkin lymphoma (cHL). A better understanding of MHC-I dysregulation in B-cell lymphomas is necessary to identify factors related to the risk, development, and progression of these tumors.

In this thesis, we investigate the role of MHC-I dysregulation in DLBCL and cHL through the application of computational approaches to study genomic data. First, we introduce some background information about the normal function of MHC-I in the immune response to cancer and viral infection as well as the phenomenon of MHC-I dysregulation in the context of cancer. We provide an overview of how factors such as germline zygosity of *HLA* class I (*HLA-I*) genes and somatic alterations in the genes *B2M* and *HLA-I* that encode the protein subunits of MHC-I contribute to the development of DLBCL and cHL.

Second, we present a study of the effects of *HLA-I* allele zygosity on survival in a cohort of 519 DLBCL patients treated with R-CHOP immunochemotherapy stratified by molecular subtype. Homozygosity in *HLA-I* was associated with a worse overall survival in patients whose tumors were classified in the "EZB" subtype, associated with somatic mutation in the epigenetic regulator *EZH2*. We find an association between the zygosity of the genes *HLA-B* and *-C* specifically and overall survival in EZB-DLBCL. These findings indicate that *HLA-I* zygosity may

be a risk factor for worse clinical prognosis in patients with the EZB subtype of DLBCL.

Third, we present a study of the genetic landscape of cHL tumors that are associated with infection with Epstein-Barr virus (EBV). We analyze inherited *HLA-I* allele types, somatic mutations, copy number changes, and mutational signatures in a cohort of 57 cHL patients (15 EBV-positive). We find that EBV-positive cHL is genetically distinct from EBV-negative cHL and is characterized by lower somatic mutation load and different activities of mutation signatures. Further, we find that cHL tumors are characterized by different patterns of MHC-I dysregulation depending on the EBV infection status. Germline homozygosity in *HLA-I* was associated with the EBV-positive subtype of cHL, while somatic alterations in *HLA-I* were associated with the EBV-negative subtype of cHL. These results suggest that inherited *HLA-I* homozygosity may be a risk factor for the EBV-positive subtype of cHL.

Fourth, we expand our study of *HLA-I* in virus-associated cHL to perform a comparative analysis of virus-positive and virus-negative tumors in nine cancers linked to five viruses. We find that virus-positive tumors occur more frequently in males and show geographical disparities in incidence. Genomic analysis of 1,658 tumors reveals virus-positive tumors exhibit distinct mutation signatures, recurrent mutations in the RNA helicases *DDX3X* and *EIF4A1*, and a lower somatic mutation burden compared to virus-negative tumors of the same cancer type. We find that germline homozygosity in *HLA-I* is a potential risk factor for the development of EBV-positive cHL, but not other common virus-associated solid or hematological malignancies.

Finally, we present in the Appendix a study of the genomic characterization of plasmablastic lymphoma (PBL), a rare EBV-associated B-cell lymphoma that occurs in the context of immunodeficiency caused by human immunodeficiency virus (HIV) infection. We find that PBL is characterized by mutations leading to constitutive activation of the JAK-STAT pathway.

We additionally identify recurrent mutations in immune-related genes, such as *B2M*. These findings indicate a potential role for MHC-I and immune dysregulation in the pathogenesis of other B-cell lymphomas.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I thank my family for their support and encouragement. I also thank my research mentor, Raul Rabadan, and the members of my thesis committee, Itsik Pe'er, Laura Pasqualucci, and Katia Basso, for their guidance and feedback. I would also like to thank my colleagues and collaborators, including Arnold Levine, Enrico Tiacci, Gianluca Schiavoni, Zhaoqi Liu, Ioan Filip, Junfei Zhao, Marco Fangazio, Erik Ladewig, and Yoonhee Nam. Thanks to the Columbia University MD/PhD Program and the Integrated Program in Cellular, Molecular, and Biomedical Studies. The studies in this thesis were funded by an NIH Medical Scientist Training Program grant (T32GM007367) and the SU2C Convergence Program.

Chapter 1: Introduction

1.1 The major histocompatibility complex class I in the adaptive immune response

The major histocompatibility complex class I (MHC-I) is a heterodimer on the surface of most nucleated cells whose function is to present endogenous peptides for recognition by cytotoxic T lymphocytes (CTLs) [1]. Its structure consists of a heavy chain molecule encoded by one of three *HLA* class I (*HLA-I*) genes and a light chain molecule, beta-2 microglobulin, encoded by the gene *B2M* (**Figure 1.1**). The heavy chain molecule consists of three extracellular domains, alpha1, alpha2, and alpha3, encoded by exons 2, 3, and 4 of the *HLA-I* gene, respectively [2]. The alpha1 and alpha2 domains form the peptide binding cleft. The alpha3 domain anchors the heavy chain to the transmembrane region and binds to the CD8 receptor on CTLs [1]. A human cell displays an estimated 100,000 peptide-loaded MHC-I molecules on its surface at a time [3, 4], which are continuously degraded and replaced to provide the immune system with an up-to-date picture of the proteins being produced by the cell [5].

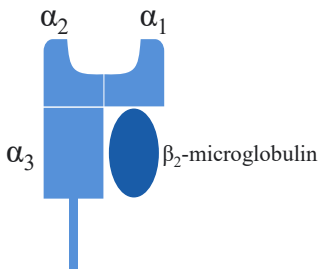


Figure 1.1: Structure of the MHC-I molecule, consisting of the heavy chain (light blue) encoded by *HLA-I*, and the light chain (dark blue) encoded by *B2M*.

Peptides are loaded onto the MHC-I in the lumen of the endoplasmic reticulum (ER) [6]. Most of the peptides presented by MHC-I derive from newly synthesized, defective ribosomal products [7] that are degraded into fragments 8-14 amino acids in length [8] by the ubiquitin-

proteasome pathway in the cytosol. Protein fragments are transported into the ER lumen by the transporter associated with antigen processing (TAP) [6]. There, fragments are bound to the peptide binding cleft of a newly synthesized MHC-I molecule, and the loaded peptide-MHC-I is secreted to the cell membrane. There are over 10^{15} unique combinations of peptides and MHC-I molecules that can be produced through this system [9].

The MHC-I is an important part of the adaptive immune response that allows CTLs to recognize and eliminate cells that display immune-stimulating proteins, or antigens. This may be seen in the response to viral infection. A cell that is infected by a virus produces viral antigens that are presented by MHC-I. An infected antigen presenting cell (APC), such as a dendritic cell, travels to a peripheral lymphoid organ, such as a lymph node, where it meets a naïve T cell [10]. The T cell is activated through binding of the T cell receptor (TCR) and CD8 co-receptor with the MHC-I of the APC, along with co-stimulation through binding of the T cell CD28 receptor to B7 protein on the APC [11]. The activated T cell then travels to the site of infection and clears other cells that display the viral antigen. This is accomplished by stimulating apoptosis through either injection of perforin and granzyme B or binding of Fas on the target cell surface [10].

Similarly, MHC-I allows CTLs to recognize cells that have undergone malignant transformation into cancer cells, through a process called immune surveillance [12]. A cancer cell can display antigens due to inappropriate expression of proteins not normally expressed in the tissue type (e.g. cancer testis antigens) or overexpression of lineage-specific proteins (e.g. MART-1/Melan-A in melanocytes and melanoma cells) [13]. While these antigens are also found in normal cells, their abnormal expression patterns in tumor cells can allow them to overcome immune self-tolerance. Inappropriately expressed high quantities of self-antigen may allow for the activation of rare CTL clones that previously escaped negative selection against self-antigens in

the thymus [14, 15]. Additionally, a cancer cell can display antigens that derive from translation of somatically mutated genes, termed neoantigens [15]. The earliest evidence for tumor-specific neoantigens came from the work of Gross and colleagues who observed that mice whose tumors were surgically resected did not develop another tumor after inoculation with the same tumor cells [16]. More recently, the widespread availability of DNA sequencing technologies has enabled the discovery of neoantigens in many types of cancer [15].

The presentation of neoantigens, enabled by MHC-I, and subsequent elimination of recognized tumor cells is an important selective pressure that drives the proliferation of tumor clones that are able to evade this immune response [17]. This is supported by the observation that MHC-I expression is decreased in 40-90% of human tumors [18]. Immune evasion is one of the most recently recognized hallmarks of cancer [19], and understanding this process is crucial to understanding cancer pathogenesis.

1.2 Heterogeneity of MHC-I in human populations

The MHC-I region is encoded on the 6p22.1-21.3 region of the genome on the telomeric end of the MHC complex (coordinates chr6:28,510,120-33,480,577 in the GRCh38 reference assembly [20]). The MHC-I region consists of three classical *HLA-I* genes (*HLA-A*, *HLA-B*, and *HLA-C*) that encode the MHC-I heavy chain [10], and two clusters of nonclassical *HLA-I* genes that perform alternative immune functions [21]. *B2M*, which encodes the MHC-I light chain, is not located in the MHC-I region but on chromosome 15 [22].

HLA-I gene alleles follow a standard four-field nomenclature (**Figure 1.2**). The first two digits of the allele name refer to the type, or allele group, usually corresponding to the serological antigen. The next two digits refer to nonsynonymous differences in amino acid sequence that produce the specific HLA protein. This is followed by the synonymous/silent substitutions and

intronic changes. In some cases, a suffix may be added with additional information about the expression of the HLA protein, if known [23].

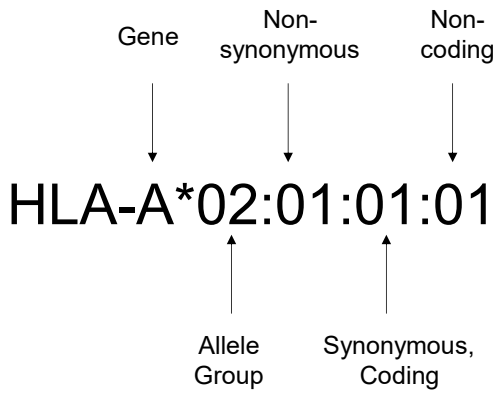


Figure 1.2: *HLA* allele nomenclature.

The class I *HLA* genes are among the most polymorphic genes in the human genome. In a study of genome-wide rates of polymorphism, all three *HLA-I* genes ranked within the top 10 most polymorphic genes normalized by gene length [24]. As of October 2022, the Immuno-Polymorphism Database / ImMunoGeneTics (IPD/IMGT) HLA Database reports 24,350 unique classical *HLA-I* alleles that have been identified in human populations [25]. The most heterogeneous *HLA-I* gene is *HLA-B*, consisting of 9,097 known alleles, followed by *HLA-A*, consisting of 7,644 alleles, and lastly *HLA-C*, consisting of 7,609 alleles. Thus, an individual typically has six *HLA-I* alleles (two each for *HLA-A*, *-B*, and *-C*) and the combination of *HLA-I* alleles is different from most other people in the population. Unlike the heavy chain *HLA-I* genes, the light chain *B2M* is highly conserved among human populations and does not display such polymorphism [26].

It has been hypothesized that the genetic diversity of *HLA-I* has evolved as a result of selective pressures from viruses. According to the MHC "heterozygote advantage" hypothesis first proposed by Doherty and Zinkernagel in 1975, individuals with two different alleles for a *HLA-I* gene will have a more extensive repertoire of antigens that can be presented on MHC-I than

patients who are homozygous in the same gene [27]. The *HLA-I* heterozygote will thus have a more robust immune response to viral infection, and heterozygosity in the *HLA-I* loci will be under positive selection. Following from this idea, Parham and Ohta in 1996 hypothesized that among *HLA-I* heterozygotes, those whose alleles encode proteins with a capacity to present a wider variety of antigens will have an advantage over those whose alleles have an overlapping peptide repertoire [28]. Some recent studies describe this as the "divergent allele hypothesis." [29] The extent of this "divergence," or physiochemical dissimilarity, may be quantified using the Grantham distance metric given the amino acid sequences of two residues i and j:

$$D = [\alpha(c_i - c_j)^2 + \beta(p_i - p_j)^2 + \gamma(v_i - v_j)^2]^{1/2}$$

where c is the composition or ratio of noncarbon to carbon elements in the side chain, p is the polarity, v is the molecular volume, and α, β, and γ are constants [30]. Evidence for MHC heterozygote advantage may be seen in the observation that genetic diversity of class I *HLA* within populations is positively correlated with the diversity of pathogens in the geographical region [31]. In the general population, the majority of people (around 80%) are heterozygous at all three *HLA-I* alleles, while around 20% of people are homozygous in at least one *HLA-I* locus [32].

1.3 Germline *HLA-I* and cancer risk[1]

HLA-I has been associated with a variety of human diseases due to its function in the adaptive immune response. Specific *HLA-I* alleles have been associated with the development of autoimmune diseases, including Graves disease (HLA-B*08 [34]), myasthenia gravis (HLA-B*08 [35]), psoriasis (HLA-C*06:02 [36]), ankylosing spondylitis (HLA-B*27 [37]), and many others.

[1] Material in this section is published in part in [33] by Marco Fangazio*, Erik Ladewig*, Karen Gomez*, Laura Garcia-Ibanez, Rahul Kumar, Julie Teruya-Feldstein, Davide Rossi, Ioan Filip, Qiang Pan-Hammarström, Giorgio Inghirami, Renzo Boldorini, German Ott, Annette M Staiger, Björn Chapuy, Gianluca Gaidano, Govind Bhagat, Katia Basso, Raul Rabadan**, Laura Pasqualucci**, and Riccardo Dalla-Favera**
*Contributed equally. **Contributed equally.

Specific links have also been described between *HLA-I* allele type and/or zygosity and response towards infections with some pathogens, including HIV, HCV, HTLV-1, dengue virus, malaria, *Leishmania*, *Plasmodium*, and *Mycobacterium* [23, 38, 39]. For instance, the alleles HLA-B*13:01:01:G and HLA-B*61 are associated with more effective infection clearance in patients with HBV [40]. HTLV-1 patients with HLA-A*02 have a decreased risk of developing HTLV-1 associated myelopathy compared to patients with other allele types [23]. HIV-1 patients who carry the *HLA-I* alleles HLA-B*35 or HLA-C*04 develop AIDS earlier than HIV-1 patients who do not have these alleles. Consistent with the heterozygote advantage hypothesis, individuals infected with HIV-1 who are homozygous in one or more *HLA-I* genes develop AIDS more rapidly and have worse overall survival compared to HIV-1-infected individuals who are fully heterozygous for *HLA-I* [41]. Overall however, fewer associations between *HLA-I* alleles and infectious diseases have been described compared to autoimmune diseases, in part due to smaller cohort sizes in the infectious disease studies [39].

The association of *HLA-I* with response to pathogens is relevant to cancer because 15-20% of cancers are directly linked to infection, the majority by viruses (termed "oncoviruses") [42-44] (**Figure 1.3**, **Table 1.1**). Thus, factors that increase susceptibility to pathogen infection, or inhibit an effective response to infection, may be expected to increase susceptibility to the associated cancer. One example in support of this hypothesis is classical Hodgkin lymphoma (cHL), a B-cell lymphoma with both EBV-positive and EBV-negative subtypes with distinct clinical and molecular characteristics [45]. HLA-A*01:01 and HLA-A*02:01 have been associated with an increased and decreased risk of developing EBV-positive cHL, respectively [46]. There is a 10-fold increase in risk of developing EBV-positive cHL in patients homozygous in HLA-A*01 compared to patients homozygous in HLA-A*02 [47].

Given the role of MHC-I in cancer immune surveillance, one might hypothesize that *HLA-I* allele type is associated with the risk of developing cancer, independent of viral infection. Allele association studies have failed to find strong associations between *HLA-I* type and risk of developing nonviral solid tumors. However, some *HLA-I* allele associations with hematological malignancies have been noted, including HLA-B*08*01 with diffuse large B-cell lymphoma (DLBCL) [48] and HLA-B*05 with classical Hodgkin lymphoma (cHL) [49]. More recently, there is emerging evidence that nonspecific *HLA-I* allele zygosity contributes to the risk of developing hematological malignancies. *HLA-I* typing of 610 non-Hodgkin lymphoma (NHL) cases and 555 controls of non-Hispanic white descent in the United States revealed that individuals who were homozygous in 2 or more *HLA-I* loci had a 1.81-fold increased risk of developing DLBCL, while those homozygous in all 3 *HLA-I* loci had a 3.66-fold increased risk of DLBCL, suggesting a heterozygote advantage in *HLA-I* for DLBCL [50]. While this study did not report whether any tumors were associated with viruses, the majority of DLBCLs would be expected to be nonviral

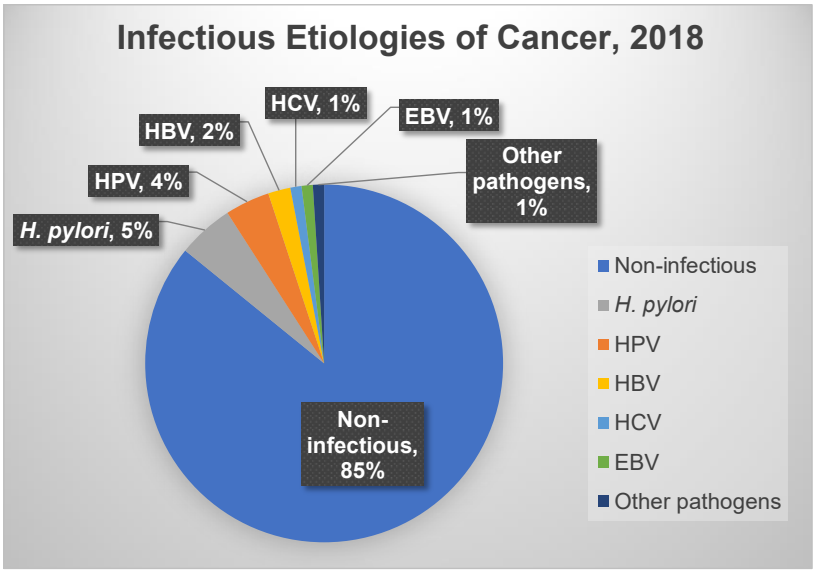


Figure 1.3: Estimated incidence of virus-associated cancers worldwide in 2018. Data obtained from de Martel et al [44].

Virus	Associated Cancer	% Caused by Virus Infection
Human papilloma virus (HPV)	Cervical Cancer (CC)	95% [51]
	Head and neck squamous cell carcinoma	30% [52]
Hepatitis B	Hepatocellular carcinoma (HCC)	56% [53]
Hepatitis C	Hepatocellular carcinoma (HCC)	20% [53]
Epstein-Barr Virus (EBV)	Hodgkin lymphoma (HL)	20-50% [45]
	Burkitt lymphoma (BL)	95% (endemic) [54] 30% (sporadic) [55]
	Plasmablastic lymphoma (PBL)	70% [56]
	Primary central nervous system lymphoma (PCNSL) (post-immunosuppression)	100% [57]
	NK T cell lymphoma (NKTCL)	100% [58]
	Gastric carcinoma (GC)	9% [59]
	Nasopharyngeal carcinoma (NPC)	100% [60]
Human herpesvirus 8 (HHV8)	Kaposi sarcoma (KS)	100% [61]
Human T-lymphotropic virus type 1 (HTLV-1)	Adult T-cell leukemia (ATL)	100% [62]
Merkel cell polyomavirus (MCPyV)	Merkel cell carcinoma (MCC)	80% [63]

Table 1.1: Human cancers linked to viral infection with the seven known oncoviruses.

in origin, as EBV contributes to <5% of DLBCLs in Western countries [64]. A follow-up study of *HLA* types in 9,922 NHL patients and 8,753 controls of European descent found that homozygosity at class I *HLA-B* and -*C* were associated with an increased odds DLBCL and marginal zone lymphoma with an OR of 1.31 and 1.45, respectively [65]. It is unknown whether *HLA-I* germline zygosity may play a similar role in Hodgkin lymphomas, either viral or non-viral in origin.

However, it is known that different lymphoma subtypes may share inherited predisposing factors. A study of 12,148 B-cell lymphoma patients in Sweden found that relatives of DLBCL patients were at increased risk of developing DLBCL or cHL but not indolent lymphomas, and relatives of cHL patients were at increased risk of developing cHL or DLBCL, but not indolent lymphomas [66].

Recently, our group investigated *HLA-I* allele type and risk of DLBCL and other tumors in the Cancer Genome Atlas (TCGA) and the UK BioBank as part of a larger study of mechanisms of *HLA-I* loss in DLBCL in collaboration with Dr. Laura Pasqualucci and Dr. Riccardo Dalla-Favera at Columbia University (**Figure 1.4**) [33]. We found that individuals with DLBCL have a greater rate of homozygosity in one or more *HLA-I* genes compared to individuals with other cancer types as well as the general population. Through analysis of UK BioBank [67] data of over 500,000 individuals from the United Kingdom, we found that there is an elevated odds ratio of DLBCL which increases with number of homozygous *HLA-I* genes: 1.11 (0.94 – 1.31, 95% CI) for one homozygous *HLA-I* gene, 1.27 (0.97-1.67) for two genes, and 1.47 (1.05-2.04) for three genes. We also saw a similar trend when comparing OR of DLBCL versus other cancers given germline homozygosity in *HLA-I*: 1.09 (0.92 – 1.29) for one homozygous *HLA-I* gene, 1.18 (0.90 – 1.56) for two genes, and 1.37 (0.98 – 1.91) for three genes. These results are consistent with the previous reports of an elevated OR of DLBCL with germline homozygosity in *HLA-I*. Additionally, we found an elevated rate of homozygosity in one or more *HLA-I* genes in individuals with cHL in the UK BioBank, suggesting *HLA-I* zygosity may be implicated in cHL as well.

Thus, an individual's inherited *HLA-I* alleles may contribute to his or her risk of developing certain subtypes of B-cell cancers. Next, we briefly describe the pathology of these tumors, which

are broadly grouped as non-Hodgkin and Hodgkin lymphomas. We then discuss evidence for somatic abrogation of MHC-I over the course of tumor development that further implicates MHC-I disruption as an important component in the pathogenesis of B-cell lymphomas.

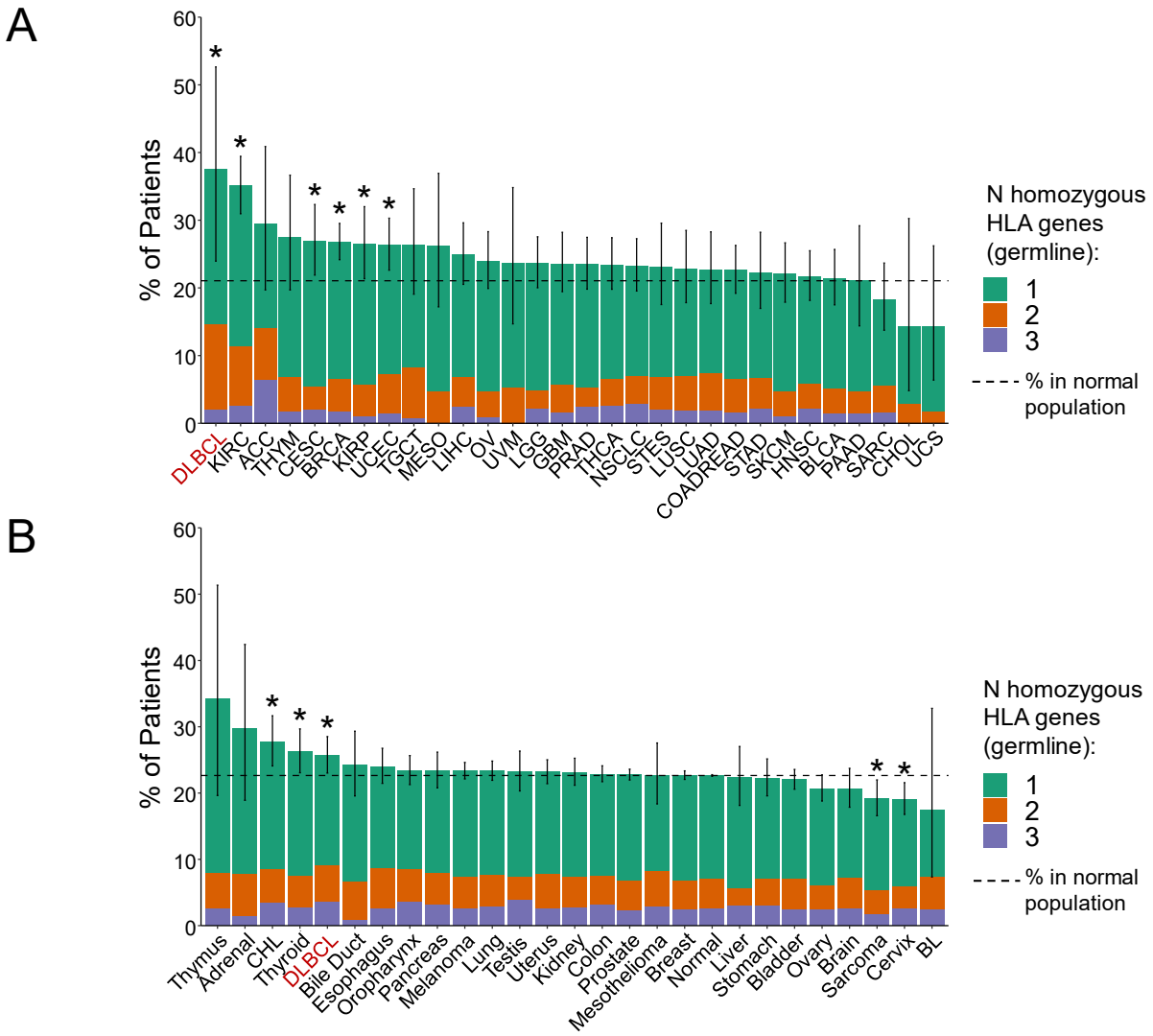


Figure 1.4. Rates of germline homozygosity in *HLA-I* genes across different cancers. The relative fraction of patients with one, two, or all three major *HLA-I* germline genes in homozygosity is shown for each of the indicated tumor types, as assessed in the TCGA (*A*) and UK Biobank (*B*) databases. Cohorts marked with an asterisk (*) had a significantly different homozygosity rate than the background normal rate calculated in healthy individuals from the UK Biobank (23%; used for comparison in the UK Biobank cohort), or from GTEx (21%; used for comparison in TCGA) ($p < 0.05$, binomial test, 95% CI shown). UK BioBank patients: Adrenal (n=64); Bile Duct (n=314); BL, Burkitt lymphoma (n=40); Bladder (n=3026); Brain (n=764); Breast (n=16245); Cervix (n=1074); CHL, Classical

Hodgkin lymphoma (n=562); Colon (n=4787); DLBCL, Diffuse large B-cell lymphoma (n=1004); Esophagus (n=1036); Kidney (n=1692); Liver (n=358); Lung (n=3332); Melanoma (n=4504); Mesothelioma (n=339); Oropharynx (n=1478); Ovary (n=1670); Pancreas (n=983); Prostate (n=10243); Sarcoma (n=850); Stomach (n=889); Testis (n=788); Thymus (n=38); Thyroid (n=708); Uterus (n=2144). TCGA patients: ACC, Adrenocortical carcinoma (n=78); BLCA, Bladder urothelial carcinoma (n=407); BRCA, Breast invasive carcinoma (n=1079); CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma (n=294); CHOL, Cholangiocarcinoma (n=35); COADREAD, Colon adenocarcinoma/ Rectum adenocarcinoma (n=562); DLBCL, Diffuse large B-cell lymphoma (n=48); GBM, Glioblastoma multiforme (n=381); HNSC, Head and neck squamous cell carcinoma (n=508); KIRC, Kidney renal clear cell carcinoma (n=507); KIRP, Kidney renal papillary cell carcinoma (n=280); LGG, Brain lower grade glioma (n=512); LIHC, Liver hepatocellular carcinoma (n=366); LUAD, Lung adenocarcinoma (n=256); LUSC, Lung squamous cell carcinoma (n=254); MESO, Mesothelioma (n=84); NSCLC, Non-small cell lung cancer (n=482); OV, Ovarian serous cystadenocarcinoma (n=418); PAAD, Pancreatic adenocarcinoma (n=128); PRAD, Prostate adenocarcinoma (n=494); SARC, Sarcoma (n=251); SKCM, Skin cutaneous melanoma (n=363); STAD, Stomach adenocarcinoma (n=225); STES, Esophagus-stomach cancers (n=203); TGCT, Testicular germ cell tumors (n=133); THCA, Thyroid carcinoma (n=499); THYM, Thymoma (n=116); UCEC, Uterine corpus endometrial carcinoma (n=528); UCS, Uterine carcinosarcoma (n=56); UVM, Uveal melanoma (n=76). Modified from [33].

1.4 MHC-I dysfunction in non-Hodgkin lymphomas

Non-Hodgkin lymphoma (NHL) is a highly heterogeneous category of lymphomas, of which 90% are of B-cell origin and 10% are of T-cell origin [68]. B-NHL includes diffuse large B-cell lymphoma (DLBCL, the most common NHL), follicular lymphoma (FL), Burkitt lymphoma (BL), mantle cell lymphoma (MCL), marginal zone lymphoma (MZL), and primary central nervous system lymphoma (PCNSL), among others. The majority of these tumors originate in the germinal center of the secondary lymphoid organs, which is the site of normal B cell activation and affinity maturation [69].

Somatic loss of expression of MHC-I has been described in several types of B-NHL. Early studies of small cohorts of NHLs described MHC-I loss in primary mediastinal lymphoma [70] and follicular lymphoma [71]. A study of DLBCL of the brain and testis found 2/15 tumor samples with loss of membrane MHC-I had inactivating mutations of *B2M* [72]. Analysis of the coding

genome of nodal DLBCL by Pasqualucci and colleagues revealed frequent inactivating mutations and deletions in *B2M* [73]. A follow-up study of 132 DLBCLs found inactivation of *B2M* by mutations and/or deletions in 29% of nodal DLBCLs [74]. These truncating *B2M* mutations were associated with lack of B2M expression, which in turn was associated with loss of HLA-I expression on the cell surface. *B2M* mutations were absent from 108 other B-NHLs analyzed the in the same study, including BL, FL, CLL, and MZL, suggesting that *B2M* truncation is unique to the pathogenesis of DLBCL. However, *B2M* mutations alone cannot completely explain the loss of expression of MHC-I that has been detected in up to 75% of *de novo* DLBCLs [74].

In our recent study of the genetic mechanisms of *HLA-I* loss in DLBCL, we found that loss of surface MHC-I was preferentially associated with DLBCL compared to other NHLs including FL and BL through immunohistochemical analysis of a panel of 657 lymphoma biopsies [33]. Analysis of tumor genetic sequencing data from 74 DLBCL samples revealed that in addition to *B2M*, *HLA-I* genes were frequent targets of somatic inactivation through truncating mutations and/or deletion. Somatic inactivation of *B2M* or *HLA-I* was found in 76.2% of DLBCLs with negative MHC-I expression on the cell surface. The mechanisms of *HLA-I* loss in the remaining fraction of MHC-I negative patients is still unknown, but may involve epigenetic silencing [75] or signals from the tumor microenvironment, such as TGF-β, which has been shown to downregulate *HLA-I* [76, 77].

DLBCL is a heterogeneous type of B-NHL that itself can be classified into different subtypes on the basis of genetic and phenotypic characteristics. The early “cell of origin” subtyping scheme described by Alizadeh and Staudt classified DLBCL into germinal center B cell-like (GCB) or activated B cell-like (ABC) on the basis of gene expression profiles [78]. ABC-DLBCL have a worse overall survival compared to GCB-DLBCL [78]. More recently, sequencing of large

cohorts of hundreds of DLBCL tumors by three independent groups (NCI [79, 80], Harvard [81], and UK-HMRN [82]) has led to the development of molecular subtyping classifications based on clustering of mutational co-occurrence. The key features of the molecular subtypes identified in these studies largely overlap, revealing 5-8 unique molecular subtypes of DLBCL [83]. In the nomenclature used by LymphGen (NCI), these include: MCD (*MYD88*^{L265P} and *CD79B* mutation), EZB (*BCL2* translocation and *EZH2* mutation), BN2 (*BCL6* fusion and *NOTCH2* mutation), ST2 (*SGK1* and *TET2* mutations), A53 (*TP53* mutation and aneuploidy), N1 (*NOTCH1* mutation), as well as genetically composite (features of more than one subtype) and unclassified cases (lacking features of any previously described subtype) [79, 80] (**Table 1.2**). The MCD and N1 subtypes are mostly ABC, and the EZB and ST2 subtypes are mostly GCB [79, 80, 83]. Mutations in MHC-I components have been described as features of specific subtypes. The LymphGen classification assigned *B2M* mutation as a feature of the A53 subtype [80], while the Harvard and UK-HMRN classifications assigned it to clusters more similar to BN2 (termed "C1" and "NOTCH2" in the

Molecular Subtype	**Genetic Features**
BN2	*NOTCH2*, *SPEN*, *BCL6* tx, *TNFAIP3*, *BCL10*, *PRKCB* mut
MCD	*MYD88*^{L265P}, *CD79A/B*, ***HLA-I,*** *PIM1*, *CDKN2A* mut
EZB	*BCL2* mut/tx, ***EZH2,*** *TNFRSF14*, *CREBBP*, *REL* mut
A53	*TP53* mut, ***B2M*** **inactivation**, aneuploidy
ST2	*SGK1*, *TET2*, *DUSP2*, *JUNB* mut
N1	*NOTCH1*, *IRF4*, *ID3* mut
Genetically Composite	Features of more than one subtype
Unclassified	Lacking features of any one subtype

Table 1.2: Genetic features of DLBCL molecular subtypes described by Wright et al. [72] (LymphGen classification scheme). Select genetic features potentially associated with MHC-I loss are indicated in bold font.

original studies, respectively) [81, 82]. The LymphGen classification assigned somatic mutations in all three classical *HLA-I* genes to the MCD subtype [80], while the Harvard classification assigned *HLA-A* to MCD and *HLA-B* to BN2 ("C5" and "C1" in the original study, respectively) [81]. Other mutated genes may modulate MHC-I expression, such as *EZH2*, which has been linked to reduced MHC-I expression on the cell surface via epigenetic silencing of *HLA-I* [75]. These findings highlight how different DLBCL subtypes are characterized by different mutations implicated in MHC-I loss.

1.5 MHC-I dysfunction in Hodgkin lymphoma

Hodgkin lymphoma (HL) is a type of B-cell lymphoma that is distinguished by the presence of Hodgkin-Reed Sternberg (HRS) cells [84]. The most common type of HL (85%) is classical Hodgkin lymphoma (cHL). These tumors consists of a few HRS cells on a background of extensive immune cell infiltrate. Around 15% of HL cases are nodular lymphocyte predominant Hodgkin lymphoma (NLPHL), which has fewer HRS cells compared to the classical type and an abundance of CD20+ lymphocyte predominant cells. Classical Hodgkin lymphoma can be further subcategorized into four subtypes based on histopathology of the tumor: nodular sclerosis (NS) (tumor nodules containing HRS on a background of fibrotic sclerosis and immune cell infiltrate), mixed cellularity (MC) (HRS cells in a mixed inflammatory cell background without sclerosis [85]), and lymphocyte-depleted or -rich [84]. Of these, nodular sclerosis and mixed cellularity are the most common subtypes, constituting 70% and 25% of cHL cases, respectively [86]. Approximately 20% of cHL cases in North America and Europe [45] and 74% of cHL cases in Africa [87] are observed in the context of Epstein-Barr virus (EBV) infection, which is associated with the mixed cellularity subtype.

Somatic dysregulation of MHC-I is frequent in HL. In a study of 361 classical Hodgkin

lymphoma samples for which MHC-I expression was assessed by immunohistochemistry staining, the rate of loss of MHC-I expression on the cell surface of cHL was 63% [88], similar to the rates of MHC-I loss seen in DLBCL. The same study found the rate of MHC-I loss was greater in EBV-negative (83.2%) compared to EBV-positive (27.4%) tumors. This is consistent with other reports that have found the rate of MHC-I loss in cHL was greater in EBV-negative (55-81%) compared to EBV-positive cHL (8-25%) [89-91]. One likely mechanism of this MHC-I loss is mutation of *B2M*, which is among the most frequent targets of somatic mutation in cHL. Mutations in *B2M* are associated preferentially with the NS subtype and are rarely seen in the MC subtype or EBV-positive cases [92, 93].

1.6 MHC-I and response to therapy in cancer

Finally, there is emerging evidence that MHC-I may be implicated in therapy response and survival in cancer. In a study of over 1,500 patients with advanced melanoma or non-small cell lung cancer, Chowell and colleagues found that homozygosity in at least one *HLA-I* locus was associated with worse overall survival in patients treated with immune checkpoint blockade (ICB) compared to patients that were fully heterozygous [94]. A follow-up study by the same group found that among patients that were fully heterozygous at *HLA-I*, those whose *HLA-I* alleles encoded proteins with a greater physiochemical difference (greater Grantham distance) had a better overall survival compared to patients whose alleles were more similar [95]. These results supports the hypothesis of a heterozygote advantage and a "divergent allele" advantage in the immune response to tumors undergoing ICB therapy. Similar results have subsequently been observed in gastrointestinal cancers [96] and kidney cancers [97] undergoing ICB therapy.

There is relatively little data available about whether a similar heterozygote advantage might be observed in the response of cancers to other kinds of treatment regimens. This question

is of particular interest in DLBCL, for which ICB therapy response rates are poor [98] and standard of care treatment remains rituximab (an anti-CD20 antibody) plus CHOP (cyclophosphamide, doxorubicin, vincristine, prednisone), termed R-CHOP, immunochemotherapy [99]. Although the addition of rituximab to CHOP chemotherapy has significantly improved DLBCL patient outcomes since its introduction two decades ago, approximately 30% of patients still fail to respond [100]. At the same time, the interaction between inherited risk factors such as *HLA-I* type and survival in DLBCL through the framework of the recently described molecular subtypes remains to be explored.

1.7 Statement of problem and organization of the thesis

In this thesis, we investigate the problem of how MHC-I dysregulation impacts the risk and development of B-cell lymphomas, with a focus on DLBCL and cHL. We develop strategies to assess MHC-I status at the germline and somatic level utilizing genetic sequencing data from DNA and RNA of these tumors. We correlate the genetic findings with clinical data including overall survival (in DLBCL) and virus infection status (in cHL).

This thesis is organized into four chapters, a conclusion, and an appendix. In Chapter 2, we explore the effect of *HLA-I* zygosity on overall survival in a cohort of DLBCL patients treated with standard of care R-CHOP immunochemotherapy. We perform *HLA-I* allele typing and survival analysis on 519 DLBCL patients, previously classified by molecular subtype on the basis of their somatic lesions. We find that *HLA-I* zygosity is associated with different overall survival trends depending on the molecular subtype of DLBCL.

In Chapter 3, we investigate the genetic characteristics of EBV-positive and EBV-negative subtypes of cHL. We analyze genetic sequencing data from 57 cHLs including 32 newly sequenced by whole genome sequencing and describe the landscape of somatic mutations detected in these

tumors. We find that EBV-positive cHL have fewer somatic mutations and different patterns of somatic mutation compared to EBV-negative cHL. Through *HLA* typing of normal DNA samples from these patients, we detect differences in the germline and somatic status of *HLA-I* depending on virus infection status in cHL. EBV-positive cases have a higher rate of germline homozygosity in *HLA-I* compared to EBV-negative cases, while EBV-negative cases have a greater frequency of somatic lesions in *HLA-I*.

In Chapter 4, we analyze common trends among nine virus-associated cancers, including cHL. We find that virus-associated cancers follow distinct epidemiological trends including a greater incidence in males and geographical variation in incidence. Analysis of DNA sequencing data reveals virus-positive tumors have a lower somatic mutation load compared to virus-negative tumors in general and are recurrently mutated in the RNA helicases *DDX3X* and *EIF4A1*. Through *HLA-I* allele typing of 1,255 patients, we find that the association between EBV-positive status and *HLA-I* germline zygosity is unique to cHL compared other virus-associated malignancies.

In the Appendix, we report (in its published format) our characterization the genomic landscape of plasmablastic lymphoma, a rare EBV-associated cancer that occurs most often in the context of immunodeficiency, such as AIDS [101]. We describe mutations in immune-regulatory genes, including *B2M*. These results indicate a potential role of MHC-I or other immune loci in the pathogenesis of some other B-cell lymphomas beyond DLBCL and cHL.

Chapter 2: *HLA* class I zygosity and overall survival in diffuse large B-cell lymphoma by molecular subtype

2.1 Introduction

The major histocompatibility complex class I (MHC-I) presents non-self antigens on the cell surface to cytotoxic CD8+ T lymphocytes as part of the adaptive immune response. Loss of expression of the surface MHC-I is a key mechanism by which tumors evade the immune system. MHC-I loss is common in diffuse large B-cell lymphoma (DLBCL), and 55-75% of DLBCL cases fail to express surface MHC-I [74]. Loss of MHC-I in DLBCL is attributed to *B2M* truncating mutations in 29% of cases [74, 88]. Somatic MHC-I loss may also be caused by mutations or copy loss of *HLA-I* or other immune genes [33]. Additionally, there is emerging evidence that the germline allele zygosity of *HLA-I* can predispose to the development of DLBCL. There is an elevated OR of DLBCL with germline homozygosity at *HLA-I* genes [33, 65], and germline homozygosity in *HLA-I* is more frequent in DLBCL compared to other cancers and the normal population [65].

DLBCL is a clinically aggressive, genetically heterogeneous B-cell lymphoma which can be classified into 5-8 molecular subtypes based on recurrent somatic mutations [79-82]. Several molecular subtypes of DLBCL are characterized by mutations in genes implicated in MHC-I loss. For instance, in the LymphGen classification scheme [80], mutations in *B2M* were attributed to the A53 subtype, while mutations in *HLA-I* were attributed to the MCD subtype. Mutations in *EZH2*, an epigenetic regulator implicated in loss of expression of MHC-I, were attributed to the EZB subtype [80]. These classifications reflect the unique mechanisms of somatic MHC-I loss among different types of DLBCL. However, it is not yet known how inherited *HLA-I* allele type is associated with DLBCL molecular subtype.

In some solid tumor patients treated with immune checkpoint blockade, *HLA-I* germline homozygosity is associated with worse overall survival [94]. Additionally, patients who are fully heterozygous at *HLA-I* whose alleles encode proteins that are more physiochemically dissimilar have a better overall survival compared to patients whose *HLA-I* alleles encode proteins that are more similar [29]. However, it is not yet known whether *HLA-I* zygosity may be associated with survival for tumors treated with other treatment modalities and/or tumors that are of hematological origin.

In this study, we determine whether zygosity of *HLA* class I alleles is associated with overall survival in a previously published [79-81] cohort of 519 DLBCL patients treated with R-CHOP immunochemotherapy. We also quantify the physiochemical distance between the encoded HLA-I proteins and determine the effect of this similarity on overall survival. Finally, we determine whether *HLA-I* zygosity is associated with age at diagnosis of DLBCL in this cohort.

2.2 Results

Homozygosity of *HLA-I* is associated with worse overall survival in EZB-DLBCL

To determine whether germline *HLA-I* zygosity is associated with overall survival in DLBCL, we performed *HLA-I* allele typing from normal DNA of 119 DLBCL patients and computed survival curves stratifying by number of germline homozygous *HLA-I* alleles (**Figure 2.1**). In the combined cohort of all molecular subtypes, we noticed a trend towards lower survival in patients with 2 or 3 germline homozygous *HLA-I* genes ($p=0.098$, log-rank test; **Figure 2.1A**). Next, we compared overall survival by molecular subtype, according to the LymphGen classifications assigned to these patients in the original study [80]. For patients of the EZB and A53 subtypes, we found that overall survival differed significantly when stratifying by *HLA-I* germline zygosity ($p=0.00088$ and $p<0.0001$, log-rank test, respectively; **Figure 2.1B,C**). In EZB

patients, overall survival was lower in patients with 2 or 3 homozygous *HLA-I* genes (HR=3.2e9[2.5e8-4.0e10], p<0.001 and HR=7.18 [1.18-43.56], p=0.032, respectively). In A53 patients, overall survival was lower in patients with 2 homozygous *HLA-I* genes (HR=120.88[1.94-7531.59], p=0.023). There was no difference in survival in other subtypes (**Figure 2.1D-E**). These results indicate that *HLA-I* allele zygosity could be linked to overall survival in certain subtypes of DLBCL.

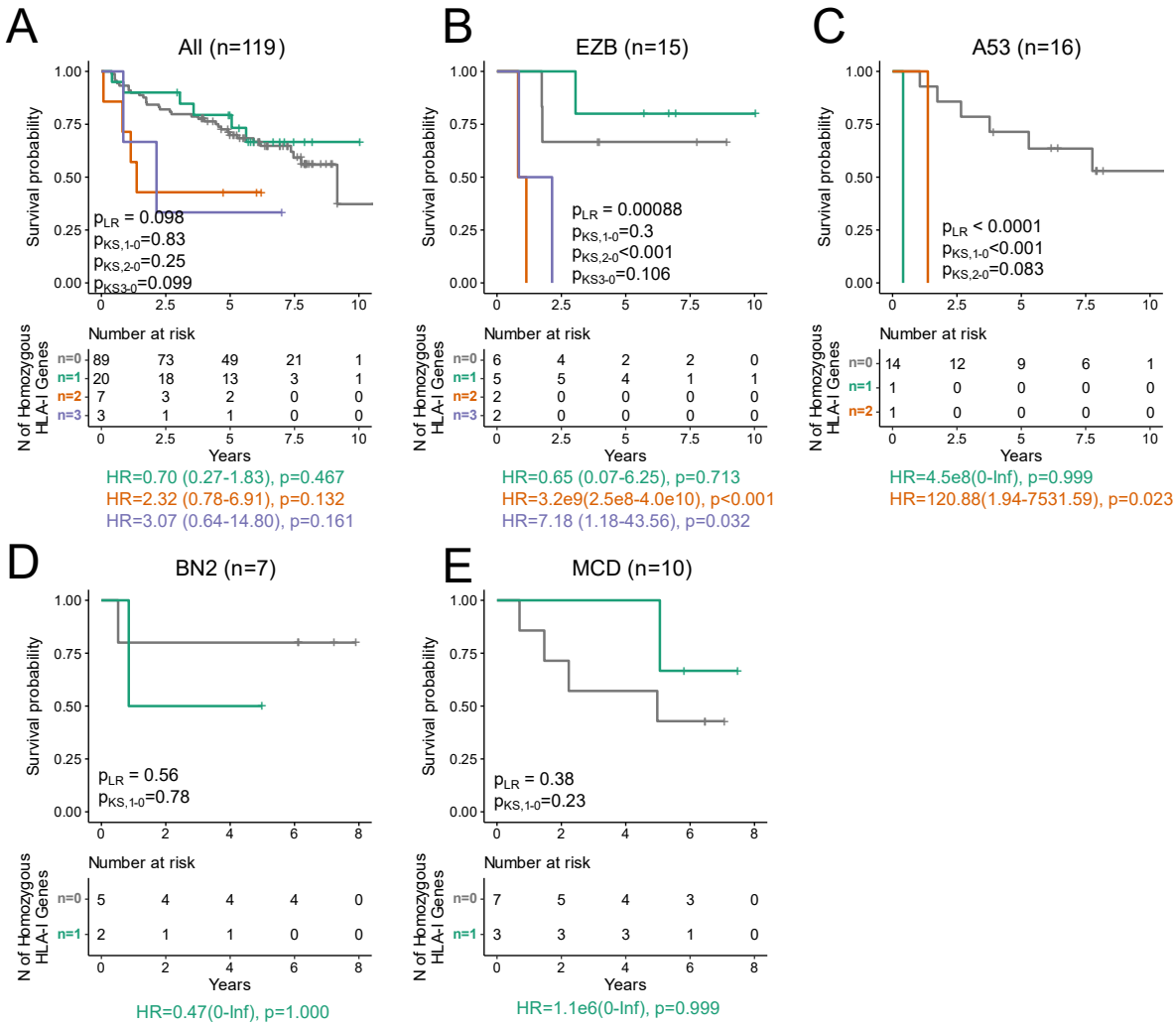


Figure 2.1: Overall survival of DLBCL patients by number of homozygous *HLA-I* genes, normal sample cohort (n=119). Overall survival is plotted for patients by molecular subtype: A) All (n=119), B) EZB (n=15), C) A53 (n=16), D) BN2 (n=7), and E) MCD (n=10). P_{LR}: p-value, log-rank test. p_{KS}: p-value, 2-sample Kolmogorov-Smirnov test. HR: Hazard ratio with chi-squared p-value.

Among the 119 patients with both tumor and normal samples available, we noted that that 94% of patients had concordant *HLA-I* zygosity classification (homozygous in at least one *HLA-I* versus fully heterozygous) in tumor and normal and 96% of patients had <= 1 allele mismatch between tumor and normal among the six inferred *HLA-I* alleles. Thus, to increase the statistical power of our study and confirm the results seen in the normal panel cohort, we performed *HLA* allele typing on 400 additional DLBCL tumor samples that lacked a matched normal and repeated the survival analysis on the combined 519 patient cohort (**Figure 2.2**). In the combined cohort of all molecular subtypes, there was no significant difference in overall survival (**Figure 2.2A**). However, we found that the overall survival of EZB-DLBCL patients differed significantly by *HLA-I* zygosity, similar to the results of the normal panel cohort (p=0.00038, log-rank test; **Figure 2.2B**). This result was consistent when analyzing data from each of the two source studies (Harvard and NCI) separately (**Figure 2.S.1**). There was no significant difference in survival in other subtypes, suggesting this association was unique to the EZB cases (**Figure 2.2C-G**). Interestingly, we noticed a trend towards better overall survival in patients homozygous in at least one *HLA-I* locus in the MCD, ST2, and N1 subtypes, though this did not reach statistical significance (p=0.08, p=0.56, and p=0.38, respectively, log-rank test; **Figure 2.2E-G**). We then computed the mean Grantham distance (or *HLA-I* evolutionary divergence, HED) between the *HLA-I* alleles of all patients to determine whether similarity between heterozygous HLA-I proteins could also contribute to overall survival. However, we found there was no difference in overall survival among patients with a HED score in the top quartile compared to patients with a lower HED score for the combined or cohort or for any subtype individually (**Figure 2.S.2**).

To further investigate *HLA-I* zygosity in the EZB subtype, we plotted survival curves of EZB patients by zygosity of *HLA-A*, *-B*, and *-C* separately (**Figure 2.3**). *HLA-A* showed a trend

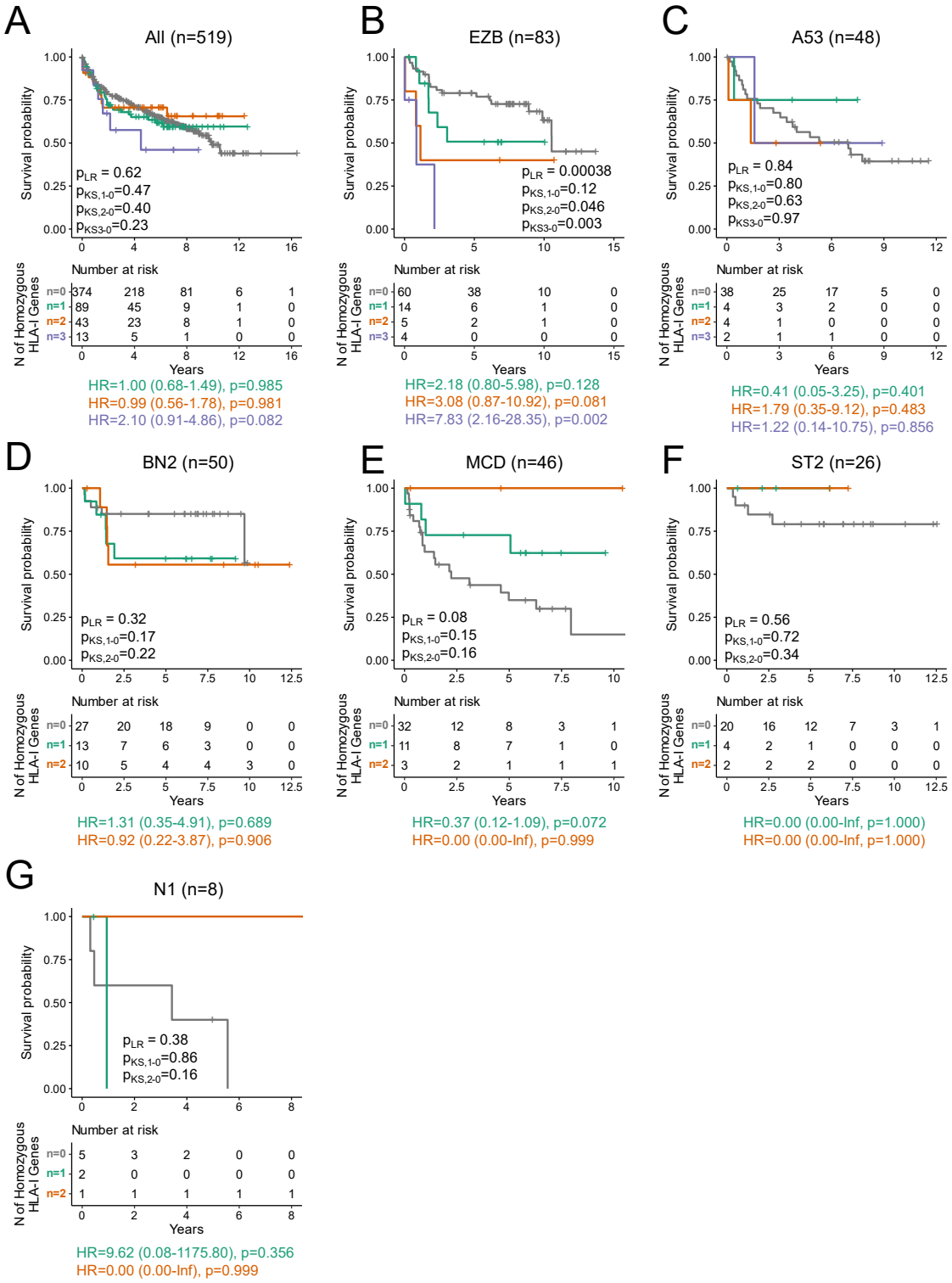


Figure 2.2: Overall survival of DLBCL patients by number of homozygous *HLA-I* genes, full sample cohort (n=519). Overall survival is plotted for patients by molecular subtype: A) All (n=519), B) EZB (n=83), C) A53 (n=48), D) BN2 (n=50), E) MCD (n=46), F) ST2 (n=26), and G) N1 (n=8). P_{LR}: p-value, log-rank test. p_{KS}: p-value, 2-sample Kolmogorov-Smirnov test. HR: Hazard ratio with chi-squared p-value.

for lower overall survival in homozygous cases, but this did not reach statistical significance (HR=1.94 (0.78-4.83), p=0.152; **Figure 2.3A**). We found that homozygosity in *HLA-B* or -*C* was associated with lower overall survival (HR=3.37 [1.39-8.19], p=0.007 and HR=6.78 [2.70-17.04], p<0.001; **Figure 2.3B-C**). Thus, the lower overall survival in *HLA-I* homozygous EZB-DLBCL patients was driven mostly by *HLA-B* and -*C*. Notably, these cases completely lacked any reported truncating somatic mutations in *B2M* or *HLA-B* and -*C*, and lacked any copy number losses in *HLA-B* or -*C*, as reported in the original studies. In contrast, 48% (11/23) of *HLA-I* homozygous EZB cases contained nonsynonymous mutations in *EZH2*, one of the most frequently mutated genes of this subtype. The lack of *B2M* or *HLA-I* lesions suggest the survival association was linked to inherited *HLA-I* allele type rather than somatic lesions for cases where *HLA-I* allele typing was determined from tumor DNA. Notably, we found that patients in this DLBCL cohort with *EZH2* mutations in the absence of MHC-I somatic mutations or deletions had a significantly lower expression of *HLA-I* genes (**Figure 2.S.3**), consistent with a previous report [75].

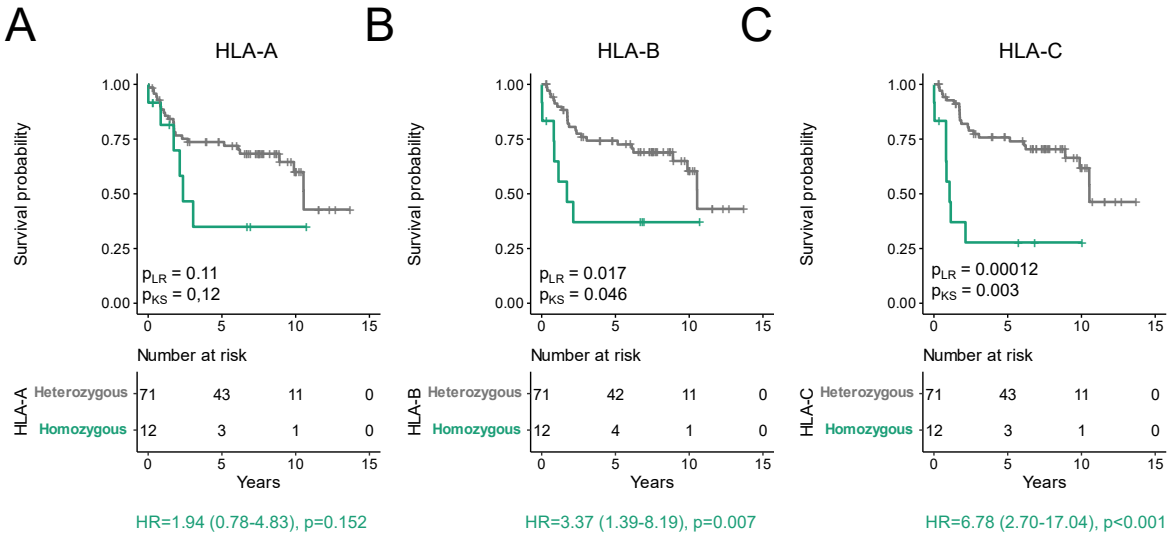


Figure 2.3: Overall survival of EZB-DLBCL patients (n=83) by zygosity of individual *HLA-I* genes: **A) *HLA-A*, B) *HLA-B*, C) *HLA-C*. P_{LR}: p-value, log-rank test. p_{KS}: p-value, 2-sample Kolmogorov-Smirnov test. HR: Hazard ratio with chi-squared p-value.**

HLA-I zygosity is not associated with age at diagnosis of DLBCL

In order to determine whether *HLA-I* homozygosity may predispose to the development of DLBCL at an earlier age, we compared age at diagnosis in patients that were homozygous in at least one *HLA-I* locus for each subtype and the cohort overall. We found there was no significant difference between age at diagnosis in patients that were fully heterozygous in *HLA-I* and patients that were homozygous in at least one *HLA-I* locus ($q > 0.1$, all comparisons; **Figure 2.S.4**). Similarly, there was no significant difference between age at diagnosis in patients that were fully heterozygous in *HLA-I* and patients that were homozygous in one, two, or three *HLA-I* genes individually ($q > 0.1$, all comparisons). Therefore, *HLA-I* allele homozygosity is not linked to earlier diagnosis of DLBCL in this cohort.

HLA-I allele type is not associated with DLBCL molecular subtype

Finally, we compared the frequency of specific alleles to determine whether individual *HLA-I* alleles were associated with increased risk of developing certain DLBCL subtypes. In total, we identified 43 unique *HLA-A* alleles, 68 unique *HLA-B* alleles, and 33 unique *HLA-C* alleles in this study cohort. There was no individual *HLA-I* allele that was more frequent in patients belonging to a specific category compared to the others ($q > 0.1$). We noticed a trend for a greater frequency of the HLA-A*01:01 allele in the ST2 subtype (58% [15/26] ST2 cases compared to 28% [138/493] of non-ST2 cases, $q=0.13$). There was also a trend for a greater frequency of the HLA-C*16*01 allele in the BN2 subtype (10% [5/50] BN2 cases compared to 1% [7/469] of non-BN2 cases, $q=0.11$). These trends indicate there may be differences in *HLA-I* allele types among DLBCL subtypes that could be made clearer in a larger cohort with a greater statistical power for detecting specific allele frequencies.

2.3 Discussion

The results of our study indicate that *HLA-I* allele zygosity may be a contributing factor in overall survival in patients with DLBCL undergoing R-CHOP immunochemotherapy. A previous study found the HLA-B44 supertype is associated with worse prognosis in patients undergoing R-CHOP therapy, suggesting inherited *HLA-I* type plays a role in tumor development and progression [102]. In a different study of 144 patients with DLBCL, there was no association between loss of MHC-I expression and response to R-CHOP therapy [103]. However, germline *HLA-I* status was not evaluated, and this study was conducted prior to the development of DLBCL molecular subtyping from DNA sequencing data, so trends within these molecular subtypes could not be determined.

The EZB subtype of DLBCL is characterized by mutations in the genes *BCL2*, *EZH2*, and *CREBBP*, among others. *EZH2* has been previously implicated in loss of surface MHC-I due to epigenetic silencing. *EZH2* encodes a protein subunit of the polycomb repressor complex 2 (PRC2) which downregulates expression of MHC-I by inhibiting NLRC5, a protein that normally activates MHC-I expression [75]. Treatment of *EZH2*-mutant human DLBCL cell lines with an EZH2 inhibitor led to increased expression of MHC-I, suggesting *EZH2* could be a potential therapeutic target [75].

Given the function of *EZH2* in epigenetic regulation of MHC-I expression, we hypothesize that the lower overall survival in *HLA-I* EZB-DLBCL homozygous cases may be due to a two-step process of MHC-I dysregulation. In the first step, inherited *HLA-I* homozygosity leads to a reduced immune repertoire so that fewer antigens can be presented by MHC-I on the cell surface. In the second step, somatic mutation in *EZH2* leads to epigenetic downregulation of MHC-I, further reducing immune repertoire and facilitating immune evasion. This is supported by a

previous study that found *EZH2* mutation was linked to decreased MHC-I expression in a murine lymphoma model [75]. Additional studies will be needed to determine whether epigenetic downregulation of MHC-I through *EZH2* activation is linked to survival in DLBCL patients.

2.4 Methods

Study cohort

The study cohort consisted of 519 DLBCL patients previously classified by molecular subtype using the LymphGen algorithm [80]. There were 294 males and 225 females and the ages ranged from 16-92. 259 patients were previously used for DLBCL molecular subtyping by Chapuy et al [81] ("Harvard" cohort) and 260 patients were previously used for DLBCL molecular subtyping by Schmitz [79] et al ("NCI" cohort). All patients were treated with R-CHOP as reported in the original source study or in the sample clinical data on GDC (https://portal.gdc.cancer.gov/).

HLA-I allele typing

HLA-I allele typing was performed from the normal sample DNA for 119 patients using PolySolver [104]. For patients with no normal sample (n=400), we performed *HLA-I* allele typing from the tumor sample DNA. To reduce the possibility of overcalling *HLA-I* homozygosity due to somatic copy loss of *HLA-I* (particularly for patients who lacked normal samples), we repeated *HLA-I* calling for all patients using a second independent *HLA-I* allele typing algorithm (Optitype [105]) and categorized each patient as homozygous at a gene only if it was homozygous in that gene in both allele callers. The sensitivity and specificity for classifying cases as homozygous in at least one *HLA-I* locus versus fully heterozygous in *HLA-I* was 100% and 94%, respectively, for tumor DNA compared to normal DNA.

Survival analysis

Survival analysis was performed using the packages "Survival" and "Survminer" in R.

Hazard ratios were computed using the package "finalfit" in R using multivariable Cox regression with covariates age, sex, cohort, cell of origin classification, and molecular subtype (when applicable). The statistical significance of differences in overall survival between two groups was assessed using the Kolmogorov-Smirnov (KS) test using the "surv2sample" package in R. The significance of differences in overall survival in three or more groups was assessed using the log-rank test using the package "Survminer" in R. Patients assigned to multiple subtypes and unclassified patients were excluded from survival analysis.

Somatic mutation and copy number analysis

Somatic mutation and copy number data for MHC-I related genes was obtained from GDC (https://gdc.cancer.gov/about-data/publications/DLBCL-2018) (NCI cohort) and Chapuy et al [81] Supplementary Table 8 (Harvard cohort). Copy number loss at *HLA-B* or -*C* was defined as any reported copy loss at the gene level for *HLA-B* or -*C* (NCI cohort) or in the 6p21.33 cytoband (Harvard cohort).

2.5 Supplementary Figures

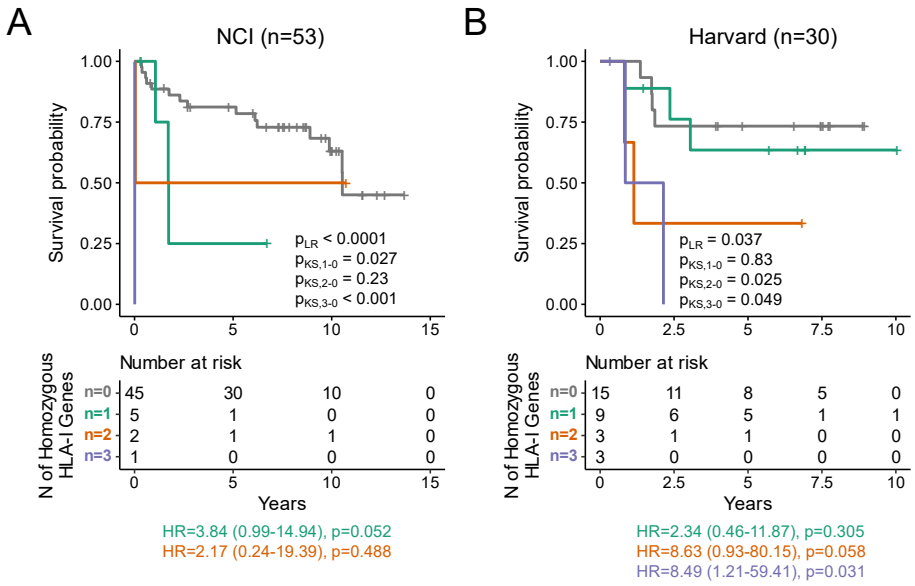


Figure 2.S.1: Overall survival of EZB-DLBCL by source study. A) Overall survival of EZB-DLBCL from [79] (NCI cohort), n=53. B) Overall survival of EZB-DLBCL from [81] (Harvard cohort), n=30. P_{LR}: p-value, log-rank test. p_{KS}: p-value, 2-sample Kolmogorov-Smirnov test. HR: Hazard ratio with chi-squared p-value.

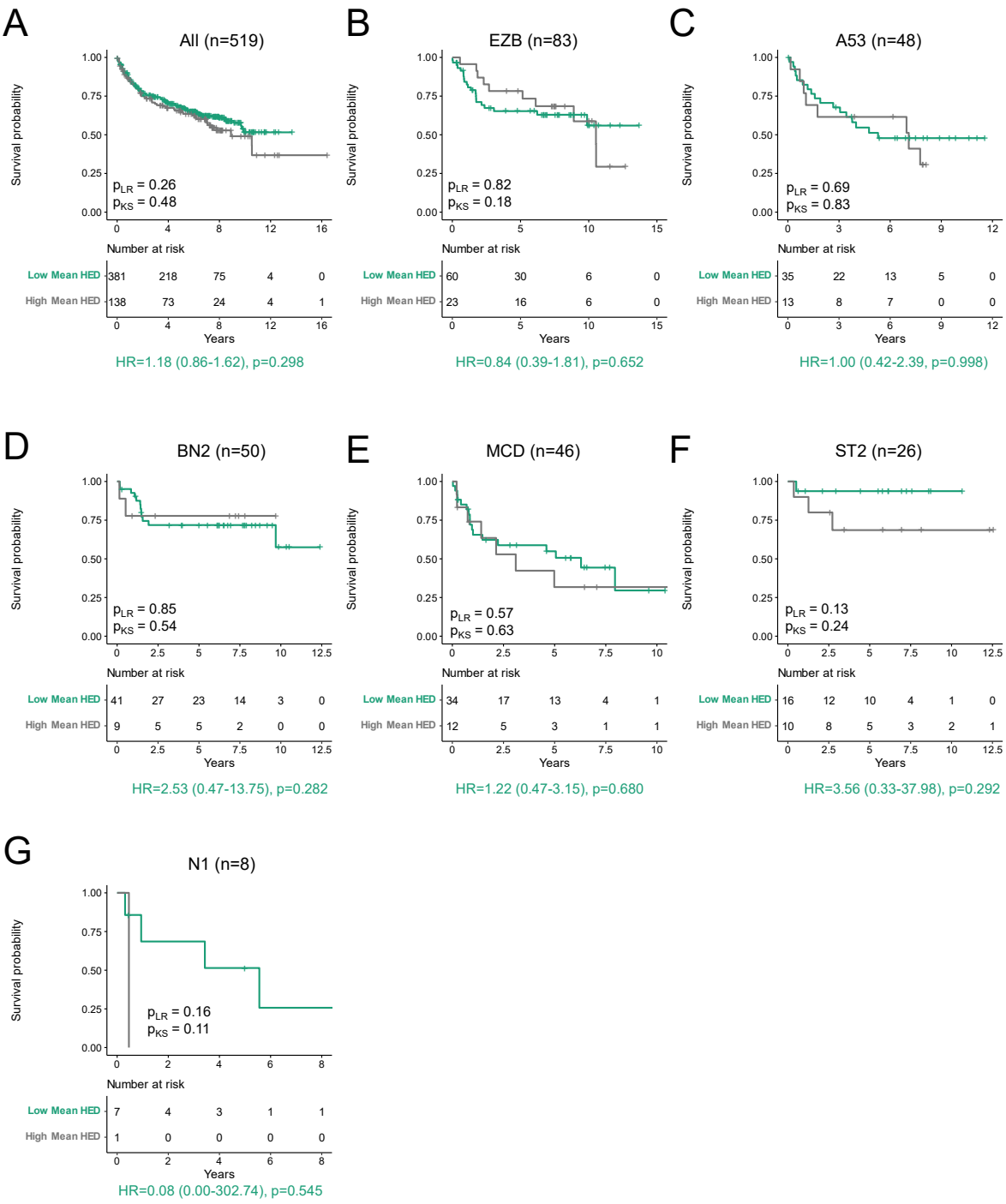


Figure 2.S.2: Overall survival of DLBCL by mean *HLA-I* evolutionary divergence (HED) score. High mean HED refers to patients in the top quartile of HED scores, while low mean HED refers to patients below the top quartile of HED scores. Overall survival is plotted for patients by molecular subtype: A) All (n=519), B) EZB (n=83), C) A53 (n=48), D) BN2 (n=50), E) MCD (n=46), F) ST2 (n=26), and N1 (n=8). P_{LR}: p-value, log-rank test. p_{KS}: p-value, 2-sample Kolmogorov-Smirnov test. HR: Hazard ratio with chi-squared p-value.

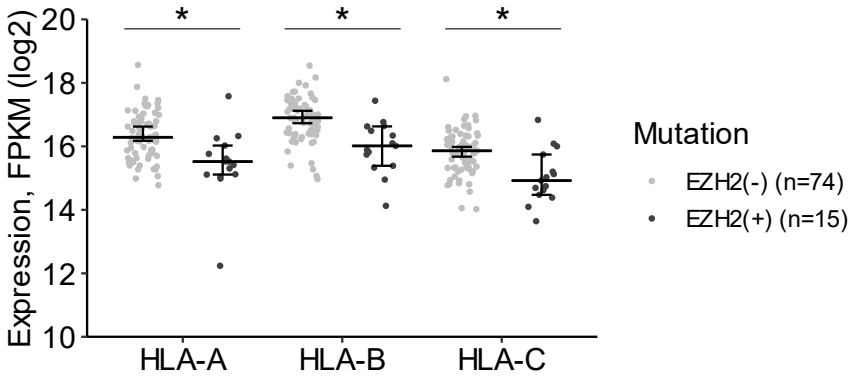


Figure 2.S.3: Expression of *HLA-I* by *EZH2* mutation status. Expression of *HLA-A*, *-B*, and *-C* in DLBCL cases with *EZH2* mutation (n=15) and cases without *EZH2* mutation (n=74) among DLBCL cases with no reported somatic mutations or copy losses in *HLA-I* or *B2M* and expression data available are shown. * p < 0.01, Mann-Whitney U test.

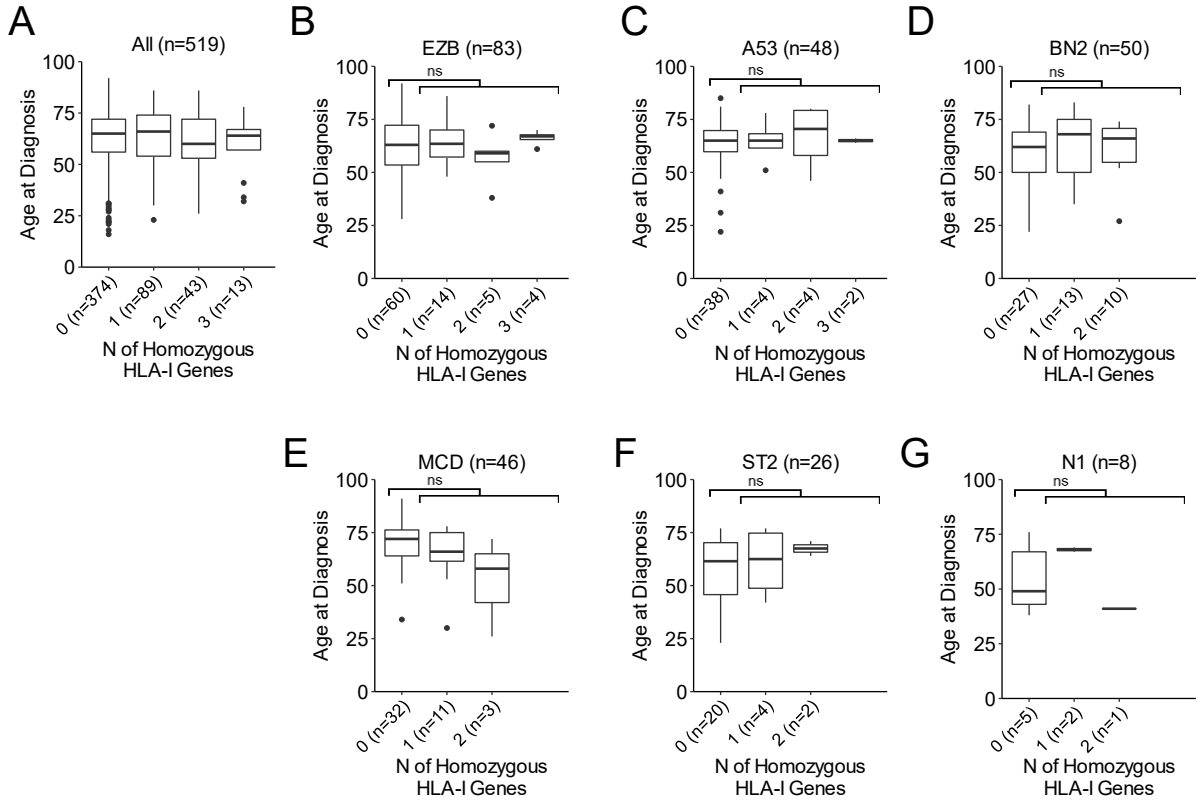


Figure 2.S.4: Age at diagnosis versus *HLA-I* zygosity in DLBCL (n=519). Ages are plotted for patients by molecular subtype: A) All (n=519), B) EZB (n=83), C) A53 (n=48), D) BN2 (n=50), E) MCD (n=46), F) ST2 (n=26), and N1 (n=8).

Chapter 3: EBV-positive and EBV-negative classical Hodgkin lymphomas are genetically distinct tumors[2]

3.1 Introduction

Classical Hodgkin lymphoma (cHL) is a common lymphoid malignancy that is characterized by the presence of multinucleated Hodgkin Reed-Sternberg (HRS) cells. Approximately 20% of cHL cases in North America and Europe are observed in the context of Epstein-Barr virus (EBV) infection [45], and these cases exhibit unique clinical and genetic characteristics. EBV-positive cHL is most strongly associated with the mixed cellularity (MC) subtype, which is characterized by HRS cells in a mixed inflammatory cell background without sclerosis [85]. Infected HRS cells usually exhibit an EBV latency II program characterized by the expression of viral proteins EBV nuclear antigen 1 (EBNA1) and latent membrane proteins 1 (LMP1) and 2a (LMP2a) [106]. EBNA1 is integral to the replication of viral DNA. LMP1 and LMP2a promote tumorigenesis by mimicking an activated CD40 receptor and B cell receptor signaling, respectively, which lead to the downstream activation of the NF-κB and PI3K-AKT pathways [106]. In EBV-positive cHL cases, nearly all HRS cells (and no normal cells) are infected, suggesting EBV infection is an early event in lymphomagenesis [106] that is transmitted clonally. However, the differences between somatic changes that occur in cHL in the context of EBV infection remain poorly understood.

Previous genomic studies [93, 107] of cHL by whole-exome sequencing (WES) suggested a lower somatic mutation burden was associated with EBV infection. However, these analyses

[2]Material in this chapter is contained wholly or in part in a manuscript in preparation by Karen Gomez*, Gianluca Schiavoni*, Yoonhee Nam, Jean-Baptiste Reynier, Cole Khamnei, Michael Aitken, Giuseppe Palmieri, Antonio Cossu, Arnold Levine, Carel van Noesel, Brunangelo Falini, Laura Pasqualucci, Enrico Tiacci**, and Raul Rabadan** *Contributed equally. **Jointly supervised this work.

included only a limited number of EBV-positive samples (4-8 cases). These previous studies were also limited to exomes, which limits what can be discovered about genome-wide patterns of mutation, such as in noncoding regions.

Here, we characterize for the first time the genomic landscape of EBV-positive and EBV-negative Hodgkin lymphoma by WGS of laser micro-dissected HRS cells from 32 patients, including 9 EBV-positive cases (**Figure 3.1A**). To identify recurrent driver lesions, we then extend the analysis to 25 additional cHL samples previously sequenced by WES by our group [93] and others [107] for a total of 57 patients (16 EBV-positive and 41 EBV-). Finally, we describe how somatic mutation load and germline zygosity at the *HLA-I* loci of cHL differ between EBV-positive and EBV-negative cases. These results paint a more complete picture of the genomic processes underlying cHL lymphomagenesis in the EBV-negative and EBV-positive subtypes.

3.2 Results

EBV-positive cHL genomes harbor fewer somatic mutations and copy number alterations compared to EBV-negative cHL

We subjected tumor and normal cells microdissected from frozen lymph node biopsies of 32 cHL cases to WGS at a median depth of 44X (IQR: 42-45X). Matched normal unamplified DNA from blood samples of 5 cHL cases was analyzed in parallel by WGS at a median depth of 41X (IQR: 40-41). Following our previous observation of a significantly lower exome mutation burden in EBV-positive compared to EBV-negative cHL cases [93], we first compared the genome-wide mutation burden (point mutations and small insertions/deletions) in EBV-positive (n=9) versus EBV-negative (n=23) cases. We found a considerably lower number of clonal somatic mutations in EBV-positive cHL than in EBV-negative cHL (median of 112, range 30-3,873, and 6,847, range 9-14,601, respectively; p= 3.2e-4, **Figure 3.1B**) despite similar coverage

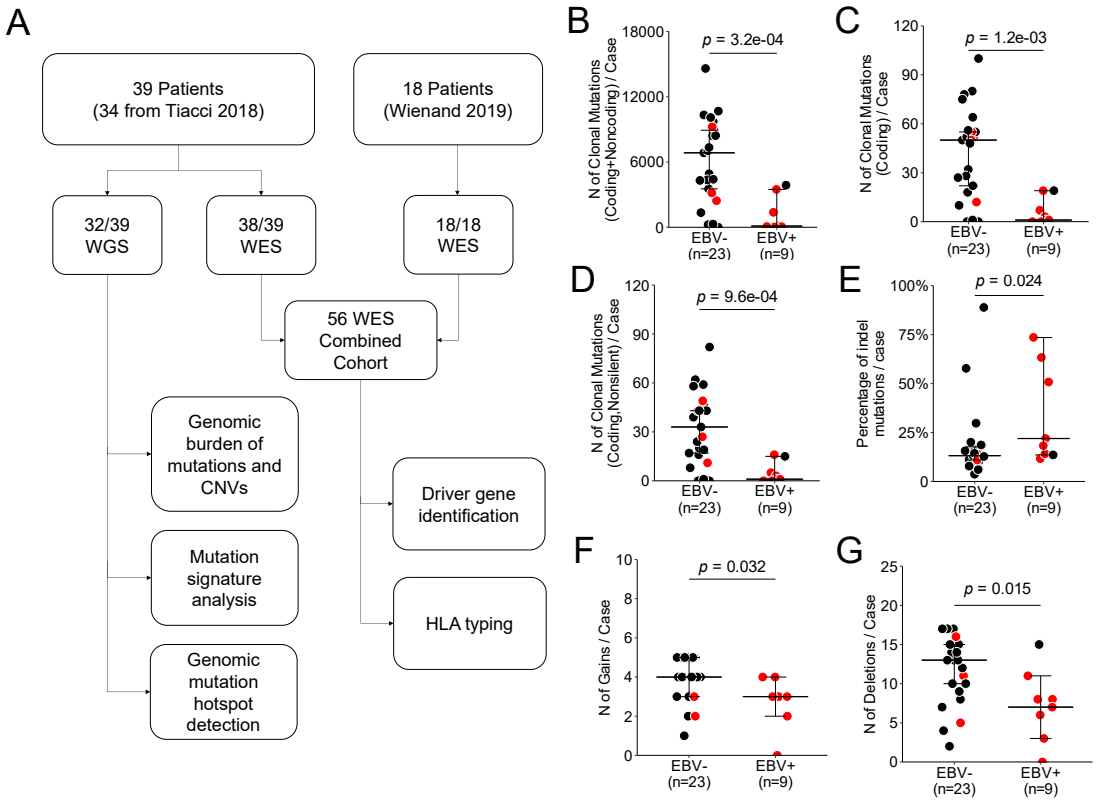


Figure 3.1: EBV-negative cHL genomes harbor a greater number of somatic mutations and copy number aberrations compared to EBV-positive cHL. A) Experimental design of the study. The mutation load for each patient in the study cohort was calculated as B) the total number of clonal mutations (point mutations and short indels), C) the total number of clonal mutations in coding regions of the genome, and D) the total number of clonal, nonsilent mutations in coding regions of the genome. E) Proportion of clonal mutations that are insertions/deletions in each case. F) Count of copy gains in 32 cHL sequenced by WGS. G) Count of copy deletions in 32 cHL sequenced by WGS. In all panels, p-values were computed with the Mann-Whitney U test, and mixed cellularity cases are colored in red.

depth between these two disease groups (both median of 44X, respectively, in the tumor samples [p=0.18], and 43X and 44X, respectively, in the normal sample [p=0.23]). The same pattern was found when restricting the analysis to coding mutations (median of 1, range 0-19, and 50, range 0-100, respectively; p=1.2e-3, **Figure 3.1C**) or to nonsilent mutations in coding regions (median of 1, range 0-16, and 33, range 0-82, respectively; p=9.6e-4, **Figure 3.1D**), consistent with previous results of mutation load by WES [93, 107]. In contrast, the relative fraction of short insertions/deletions was greater in the EBV-positive versus EBV-negative cases, suggesting

potentially different underlying mechanisms (median of 22% vs 13%, respectively; p=0.024, **Figure 3.1E**). These data cannot be explained by age-associated processes, as patients with EBV-positive cHL presented with significantly older age compared to the latter group (median of 63 and 33 years, respectively [p=0.029]) (**Figure 3.S.1**).

As expected, most EBV-positive cases were of the mixed cellularity subtype (8/9), but EBV infection was still associated with a lower clonal mutation load when restricting the analysis to this histological subtype (p=0.01, **Figure 3.1B**). Although the number of mixed cellularity EBV-negative cases sequenced by WGS was low (n=3), this result is consistent with previous observations in the coding genome of EBV-positive versus EBV-negative mixed cellularity cases [93].

To validate these findings, we used an extended cohort of 56 cHL cases (15 EBV-positive and 41 EBV-negative; including 25 cases distinct from the 32 described above), which were subjected to WES (median depth 146X in both the EBV-positive and EBV-negative groups) and comprised 38 cases from our group (34 previously published [93] and 4 newly generated; 8/38 EBV-positive; 31/38 also subjected to WGS) plus 18 cases available for download from ref. [107] (7/18 EBV-positive). All of these 56 cases were homogenously subjected to the same bioinformatics pipelines (Methods). Compared to EBV-negative cHL, EBV-positive cHL was confirmed to have a lower exome-wide load of somatic mutations (with a similar trend when restricting the analysis to the 16 mixed cellularity cases, 10 EBV-positive and 6 EBV-negative; p-value 0.11 for total clonal mutations) (**Figure 3.S.2**).

In addition to somatic mutations, we detected recurrent somatic copy number alterations with GISTIC2.0 (see Methods). The number of recurrent somatic copy gains as detected from WGS was also significantly lower in EBV-positive (median 3, range 0-4) compared to EBV-

negative cases (median 4, range 1-5; p=0.032, **Figure 3.1F**). Copy number losses were also more frequent in EBV-negative (median 13, range 2-17) cases than EBV-positive (median 7, range 0-15; p=0.015, **Figure 3.1G**).

Nonsilent alterations in the JAK-STAT and NF-κB pathways are more frequent in EBV-negative cHL than EBV-positive cHL

We compared the mutation frequency of key Hodgkin lymphoma genes in EBV-negative and EBV-positive cHL from WES of 56 cHL tumors (**Figure 3.2A**). Consistent with previous studies [108, 109], we found the JAK-STAT pathway members *STAT6* and *SOCS1* were the most frequently mutated genes, being affected in 24/56 cases (43%), the majority EBV-negative (22/24 [92%]). In particular, *STAT6* missense mutations of the DNA binding domain were observed in 14/41 EBV-negative cases (34%) but in only 1/15 EBV-positive cases (7%; p=0.047, MWU test) (**Figure 3.2B,C**). Furthermore, gains or amplifications of 9p24.1/*JAK2* were significantly enriched in EBV-negative cases (31/41 [76%], versus 6/15 [40%] EBV-positive cases; p=0.024). There were also mutations in the JAK-STAT gene *CSF2RB* in 4 cases, 3/4 of which were EBV-negative. Overall, at least one of these four JAK-STAT pathway genes was targeted by genetic lesions in 35/41 EBV-negative tumors (85%) but only 7/15 EBV-positive tumors (47%; p=0.0057). Another frequently targeted pathway was NF-κB. Recurrent mutations and/or deletions of *TNFAIP3* at 6q23.3, mutations of *NFKBIE*, and/or gain/amplification of *REL* at 2p16.1 were observed in 42/56 cases overall (75%) and more often in EBV-negative cases (35/41 [85%]) than EBV-positive cases (7/15 [47%]; p=0.0057). A third pathway more often mutated in EBV-negative tumors was PI3K-AKT (14/41 [34%] EBV-negative cases versus 1/15 [7%] EBV-positive cases; p=0.047), including the pathway inhibitor genes *GNA13* and *ITPKB.* In particular, at least one nonsilent mutation in *GNA13* (missense or truncating events distributed throughout the coding sequence) or out-of-frame

tandem duplication event was detected in 10/41 EBV-negative cases (24%), while none of the 15 EBV-positive cases had *GNA13* alterations (p=0.048). The lack of *GNA13* nonsilent mutations in EBV-positive cHL was further confirmed by targeted Sanger sequencing of an additional 18 EBV-positive cases (i.e., 33 in total) (p=0.0035) (**Figure 3.2D,E**). Finally, mutations in MHC-I genes (*B2M* and *HLA* class I [*HLA-I*]) were more frequent in EBV-negative (23/41 [56%]) compared to

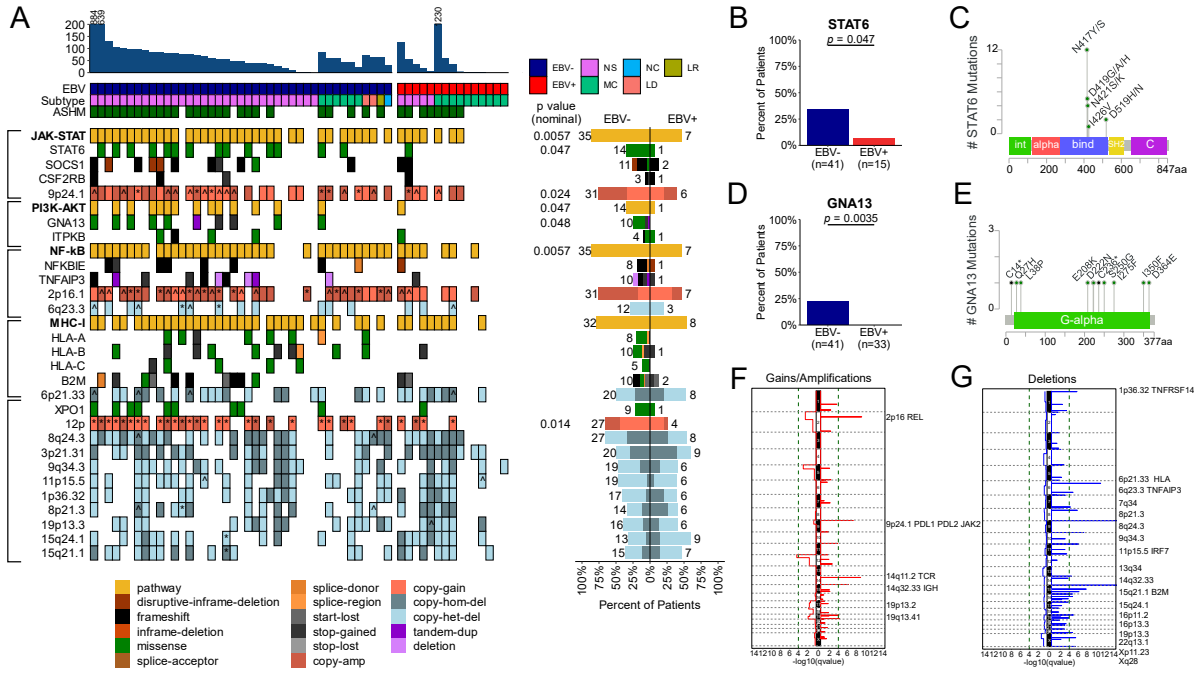


Figure 3.2: Genetic lesions in EBV-positive and EBV-negative cHL sequenced by WES. A) Genomic landscape of cHL (n=56). Clonal mutations in genes mutated in at least four patients and known to be implicated in cHL pathogenesis, as well as significant copy number peaks from GISTIC, are shown. Mutations in *HLA-I* genes were obtained from PolySolver, while other mutations were obtained from SAVI +/- Sanger sequencing. ^chromosomal copy number alteration; *arm-level copy number alteration. ASHM: at least one mutation in an aberrant somatic hypermutation target region. JAK-STAT: at least one mutation in *STAT6*, *SOCS1*, *CSF2RB*, and/or copy number alteration (CNA) in 9p24.1. PI3K-AKT: at least one mutation in *GNA13* or *ITPKB*. NF-κB: at least one mutation in *NFKBIE*, *TNFAIP3*, and/or CNA in 6q23.3 or 2p16.1. P value: Fisher's exact test. B) Counts of patients with *STAT6* nonsilent mutations in EBV-negative (n=41) and EBV-positive (n=15) cHL. P value: Fisher's exact test. C) Nonsilent mutations in *STAT6* observed in 56 cHL patients. D) Counts of patients with *GNA13* nonsilent mutations in EBV-negative (n=41) and EBV-positive (n=33) cHL. P value: Fisher's exact test. E) Nonsilent mutations in *GNA13* observed in 56 cHL patients. F) Recurrent copy number gains in cHL, identified by GISTIC (n=56). G) Recurrent copy number losses in cHL, identified by GISTIC (n=56).

EBV-positive (3/15 [20%]) cases (p=0.032), and a similar trend was observed when also including deletion in the MHC region at 6p21.33 (32/41 [78%] EBV-negative cases versus 8/15 [53%] EBV-positive cases; p=0.097). These results indicate that multiple loci known to be commonly mutated in cHL are mutated frequently in the EBV-negative subtype but rarely in the EBV-positive subtype.

APOBEC signature contributes to a greater fraction of mutations in EBV-negative than EBV-positive cHL

Single base, double base, and indel mutation signatures were detected in the 32 cHL whole genomes using a supervised non-negative matrix factorization approach informed by *de novo* signature calling (see Methods). We detected 10 signatures through *de novo* signature calling (**Figure 3.3A**), and ran supervised analysis on the 10 most similar COSMIC v3 mutation signatures based on cosine similarity score (**Figure 3.3B**). When comparing the mutation counts in each sample (**Figure 3.S.3**), COSMIC mutation signatures SBS2 and SBS13, attributed to apolipoprotein B mRNA-editing enzyme, catalytic polypeptide (APOBEC), contributed to significantly more mutations in the EBV-negative compared to EBV-positive cases (q=0.011 and q=0.0088, respectively). The signature SBS9, associated with somatic hypermutation (SHM), was also the source of more mutations in the EBV-negative compared to EBV-positive (q=0.0069), along with SBS85, associated with indirect effects of AID hypermutation (q=0.0023). The signature SBS5, whose mutation load is linearly associated with patient age [110], was more frequent in EBV-negative than EBV-positive samples (q=4.9e-4). The double base signatures DBS1 and DBS7 were detected at a higher counts in the EBV-negative than EBV-positive (q=0.11 and q=0.0023, respectively), though the overall counts of double base signatures were low (<=51 DBS mutations/patient in 30/32 patients). Finally, both of the detected indel mutation signatures

(ID2 and ID8) contributed to a higher number of mutations in EBV-negative than EBV-positive cases (q=0.0023 and q=0.0023, respectively). These results indicate that the overall higher mutation load in EBV-negative cases compared to EBV-positive is due to the accumulation of more mutations from all signatures detected in this cohort and is not due to any one specific mutagenic process.

In addition, the fraction of mutations attributed to each signature in each tumor varied based on EBV infection status (**Figure 3.3**). Specifically, the APOBEC mutation signature (SBS2) and the double-base signature associated with exposure to UV light (DBS1) contributed to a greater fraction of mutations in the EBV-negative compared to the EBV-positive samples (q=0.11 and 0.21 respectively). The overall fraction of mutations attributed to indel signature ID2, reflecting slippage during DNA replication of the template DNA strand, was also higher in the EBV-negative compared to the EBV-positive samples (q=0.21). Conversely, ID8, which is hypothesized to be caused by repair of DNA double strand breaks by non-homologous DNA end-joining mechanisms, was higher in the EBV-positive compared to the EBV-negative samples (q=0.21). These results indicate that the relative activities of mutagenic processes in cHL differ depending on EBV infection status, with APOBEC contributing to a relatively larger proportion of mutations in the EBV-negative subtype.

Aberrant somatic hypermutation associated hot spots are mutated more frequently in EBV-negative than EBV-positive cHL

The process of somatic hypermutation has been shown to function aberrantly in 55% of cHLs [111]. To compare the role of aberrant somatic hypermutation (ASHM) between EBV-negative and EBV-positive cHL, we compared the mutation load in ASHM-associated regions in whole-genome sequenced cases by subtype (Methods). We found EBV-negative cases had a

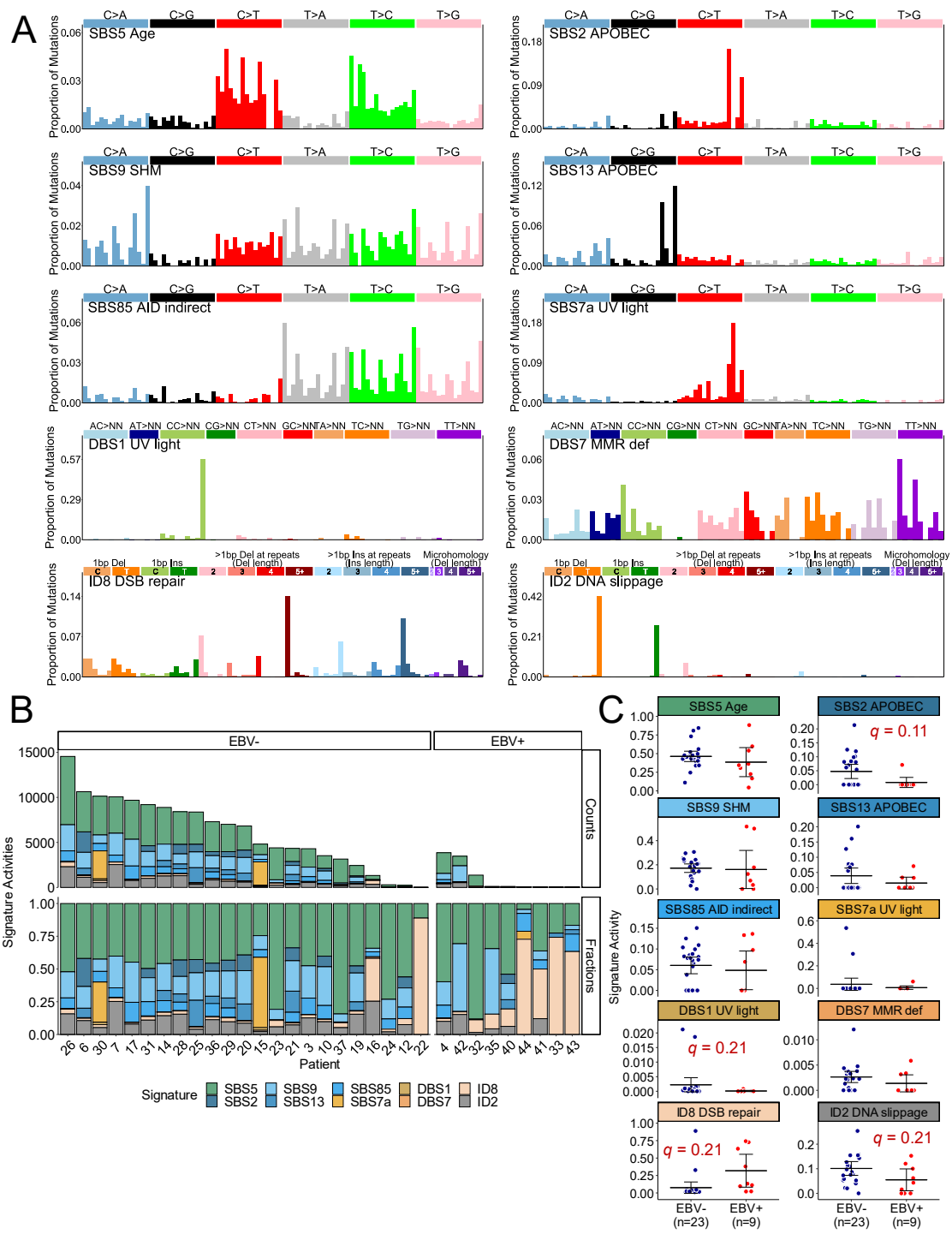


Figure 3.3: APOBEC signature contributes to a greater proportion of total mutations in EBV-negative than EBV-positive cHL. A) Mutation signatures identified in 32 cHL patients by *de novo* signature analysis with Palimpsest. Signatures are annotated with the most similar COSMIC v3 mutation signature based on cosine similarity score. B) Frequency and counts of mutation signatures in cHL patients determined by supervised mutation signature analysis with Palimpsest. C) Proportions of mutations attributed to each mutation signature in EBV-negative versus EBV-positive cHL (p values, Mann Whitney U test, BH corrected).

greater number of mutations in ASHM-associated regions overall (**Figure 3.4A**). There were also a greater number of uniquely mutated ASHM-target genes (**Figure 3.4B**) and a greater number of mean mutations per ASHM region in the EBV-negative than the EBV-positive patients (**Figure 3.4C**). These patterns reflect both the overall greater mutation load in these samples as well as the relatively larger proportion of mutations that are attributed to ASHM in EBV-negative cases.

In order to further elucidate mutational processes in the cHL genome, we identified hot spot regions of somatic mutation through application of Gamma-Poisson regression on genomic intervals using the R package Fishhook [112] (Methods). Eight somatic mutation hot spots (six SNV, two indel) were identified, all of which were in noncoding regions (**Figure 3.4D-F**). Two hot spots (one SNV, one indel) were within 2 kb of the transcription start site in *IGLL5*, a known target region of ASHM. Another hot spot was 196 kb downstream of *BCL6*, which is a gene usually targeted by physiologic SHM in normal GC B cells [113, 114] but also enriched in tumor-associated events [115], though the hot spot did not lie within the previously reported 2-kb region downstream the transcription start site. Another recurrent mutation hotspot was in intron 1 of the gene *AICDA*, which encodes the activation-induced cytodine deaminase that generates mutations during the process of SHM in B cells. 27% of SNVs in these hot spots occurred at RGYW motifs, frequently targeted by SHM, compared to 10% of SNVs in other regions of the genome (p=6.6e-12). Notably, all of these mutation hot spots were found overwhelmingly in the EBV-negative cases, and were virtually unmutated in EBV-positive cases (0-1 case each in all eight hotspots identified). These results further underline differences between mutagenic processes in these two groups of patients and indicate a potential role of ASHM in EBV-negative cHL specifically.

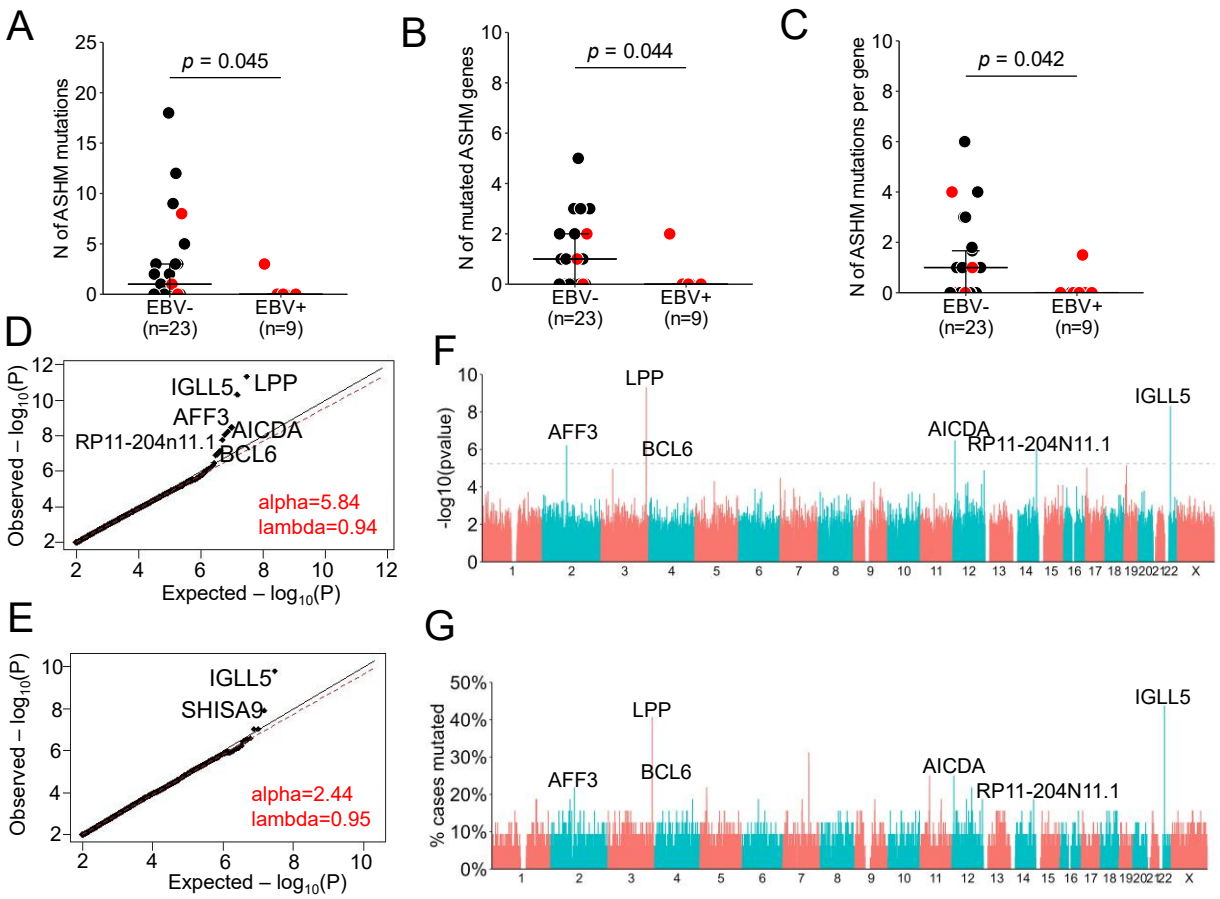


Figure 3.4: Aberrant somatic hypermutation-associated regions are mutated more frequently in EBV-negative than EBV-positive cHL. A) The number of mutations within 2 kb of the transcription start site of 128 ASHM genes per patient. B) The number of unique genes from (A) per patient. C) The mean number of mutations within 2 kb of the transcription start site of 128 ASHM genes per gene per patient. Red: MC subtype. D) Q-Q plot of SNV hot spots in cHL, annotated by gene for q < 0.25. E) Q-Q plot of indel hot spots in cHL, annotated by gene for q < 0.25. F) Significance of peaks of SNV hot spots from (D). G) Percentage of cHL cases mutated at SNV hot spots from (D).

Germline homozygosity in *HLA* class I is more frequent in EBV-positive than EBV-negative cHL

The *HLA-I* loci have been implicated in the pathogenesis of cHL, and have also been associated with other EBV-associated tumors such as nasopharyngeal carcinoma [116]. Thus, we assessed the *HLA-I* status in our cohort of EBV-positive and EBV-negative cHL cases. The count of total mutations and nonsynonymous mutations in *HLA-I* per patient was greater in EBV-

negative than EBV-positive cases (**Figures 3.5A, 3.S.4**), consistent with the trend for greater somatic mutation load in EBV-negative cases seen genome-wide (**Figure 3.1**). Missense or truncating mutations in one or more *HLA-I* loci were detected in 17/56 cases, including 8/56 in *HLA-A*, 11/56 in *HLA-B*, and 5/56 in *HLA-C* (**Figure 3.3A**). For all *HLA-I* loci, the majority of patients with missense or truncating mutations were EBV-negative: 8/8 for *HLA-A* (p=0.093), 10/11 for *HLA-B* (p=0.25), and 6/6 for *HLA-C* (p=0.31, Fisher's exact test), collectively reaching statistical significance despite the small number of cases (n=16/41 [39%] EBV-negative cases versus 1/15 [7%] EBV-positive cases; p=0.023). 2/56 cases (both EBV-negative) had biallelic nonsynonymous mutations in at least one *HLA-I* gene, and 1 case had biallelic truncating mutations in *HLA-B*. An additional 12/56 cases harbored missense or truncating mutations in *B2M* (detected through WES and/or Sanger sequencing [93]), and 3/56 cases (3/3 EBV-negative) harbored a missense or truncating mutation in both *B2M* and at least one *HLA-I* gene. In total, 26/56 (46%) cases were mutated in *HLA-I* or *B2M*, the majority of which (23/26, 88%) were EBV-negative (p=0.032; **Figure 3.3A**). Similarly, somatic loss of heterozygosity caused by allele deletion tended to occur more often in EBV-negative than EBV-positive cases (11/11 [100%] versus 5/7 [71%] evaluable cases, p=0.14, MWU test) (**Figure 3.5B**). Overall, EBV-negative cHL more frequently carried somatic lesions potentially disturbing MHC-I presentation of tumor neoantigens, which in turn are likely more numerous in this group than in EBV-positive cHL, due to higher somatic mutation burden (analogous to carcinogen induced-tumors such as lung carcinoma and melanoma) [107].

We performed germline allele typing of the same samples using the PolySolver [104] algorithm, as emerging evidence points to frequent homozygosity in patients with B-cell lymphomas [33, 65] as a possible means to escape immune evasion. We found the HLA-A*02*01

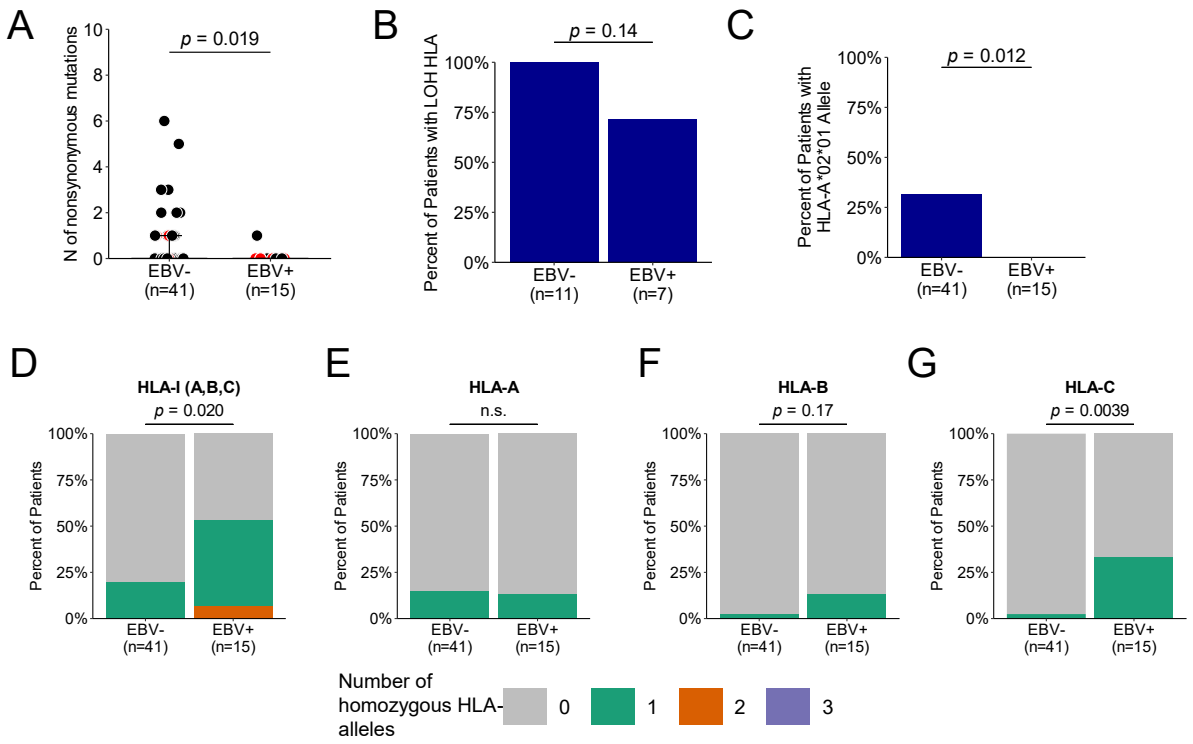


Figure 3.5: Class I *HLA* in 56 cHL sequenced by WES. A) Count of nonsynonymous mutations in *HLA-I* genes in EBV-negative versus EBV-positive cHL (p value, Mann-Whitney U test). B) Percent of patients with loss of heterozygosity of *HLA-I* by EBV status (n=18) (p value, Mann-Whitney U test). C) Percent of patients with HLA-A*02*01 allele by EBV status. D) Percent of EBV-positive and EBV-negative patients with 0, 1, 2, or 3 sets of homozygous *HLA-I* alleles (p value, Fisher's exact test). E-G) Percent of EBV-positive and EBV-negative patients with homozygosity in *HLA-A*, *-B*, and *-C* (p values, Fisher's exact test).

allele was more frequent in the EBV-negative than EBV-positive cases (p=0.012), which is consistent with previously described differences in *HLA-A* allele types between these two disease subgroups [46] (**Figure 3.5C**). Interestingly, when we compared rates of homozygosity at one or more *HLA-I* genes in the germline (**Figure 3.5D**), EBV-positive patients were significantly more likely to be homozygous in at least one of the three *HLA-I* loci compared to EBV-negative (p=0.020). Specifically, EBV-positive patients were significantly more likely to be homozygous in *HLA-C* than EBV-negative patients (p=0.0039), and a similar trend was observed in the rate of homozygosity in *HLA-B* (**Figure 3.5E-G**). These results support previous work which has

identified increased rates of germline homozygosity in cHL compared to other cancers [33].

3.3 Discussion

The first finding of our study is that EBV-positive classical Hodgkin lymphoma is affected by fewer somatic lesions genome-wide compared to cases without EBV infection. This pattern was observed consistently for somatic SNVs, short indels, and copy number aberrations, and points to the unique role of the Epstein-Bar viral machinery in the tumorigenesis of EBV-positive cHL. This finding supports the hypothesis that viral proteins drive oncogenesis without the need for the kinds of somatic alterations observed in the EBV-negative subtype. This is achieved through the expression of the latent membrane proteins LMP1 and LMP2a in EBV-infected HRS cells. LMP1 mimics CD40 and activates the NF-κB pathway, while LMP2a mimics B cell receptor signaling and can activate the PI3K-AKT pathway [117].

Our study confirms earlier WES analysis of cHL showing frequent mutations in the JAK-STAT gene *STAT6* and the PI3K inhibitor [118] gene *GNA13* [93] and provides robust evidence for the association of these mutations with EBV-negative, but not EBV-positive cHL. The absence of somatic lesions in *GNA13* in EBV-positive cases may be explained by the function of the latent membrane protein LMP2a produced in infected cells. LMP2a activates the PI3K-AKT pathway, which may eliminate the selective pressure for PI3K-AKT-activating mutations in the tumor cells.

Previous studies [107, 119] of mutation signatures in cHL have identified a mixed signature of spontaneous deamination at CpGs attributed to aging. We identified a mixed signature characterized by 1) C>T mutations at NCG trinucleotides, similar to the previously described aging signature and COSMIC1, as well as 2) frequent T>C mutations. This signature was most similar to SBS5 from COSMIC-v3, which has been associated with cellular turnover and aging. This signature was not correlated with patient age in our data set, consistent with the previous report,

strengthening the hypothesis that CpG deamination is driven by an early increase in the division rate of HRS cells [107].

We also identified signatures of the apolipoprotein B mRNA-editing enzyme, catalytic polypeptide (APOBEC) family of cytidine deaminases in our cohort, previously described in cHL [107, 119]. The signature activity of APOBEC (SBS2) was more active in EBV-negative than EBV-positive samples, indicating this mutagenic process plays a larger role in the tumorigenesis of EBV-negative than EBV-positive cHL. In addition to signatures of APOBEC, we identified signatures associated with the process of somatic hypermutation (SBS9, SBS85). A previous study of mutations in four regions targeted by ASHM determined by direct DNA sequencing found that 55% of cHLs were mutated in one or more ASHM target genes, and 30% harbored mutations in two or more genes [111]. Here, we found a recurrent mutation hotspot for both SNVs and indels occurred in a region of *IGLL5* targeted by ASHM, and was mutated in EBV-negative cases nearly exclusively. Mutation hotspots located in *AICDA* and downstream of the gene *BCL6* were also identified overwhelmingly in the EBV-negative cases. Our findings indicate ASHM may contribute to the mutational processes in EBV-negative more often than EBV-positive cHL.

30% of samples in this cohort contained at least one missense or truncating mutation in an *HLA-I* gene. This follows reports by our [93] and other groups [107] of recurrent mutations that impair function of the MHC-I complex in cHL, including *HLA-B* and related loci such as *B2M*. Almost all cases with a missense or truncating mutation in *HLA-I* were EBV-negative, which is consistent with the observation that loss of MHC-I expression is more common in EBV-negative than EBV-positive cHL. An interesting new finding from our analysis is the association between homozygosity in one or more *HLA-I* loci and positive EBV infection status. cHL has been shown to have higher rates of germline homozygosity in *HLA-I* compared to other cancers [33]. This may

reflect how similarity of *HLA-I* alleles in the germline predisposes to viral infection, which in turn leads to increased risk of tumorigenesis.

This comparative analysis of the genomic landscape of cHL in the presence and absence of EBV infection reveals that EBV-negative cHL is characterized by distinct genetic features, including greater activity of APOBEC signature, recurrent mutations in *STAT6* and *GNA13*, and copy number gains at the 9p24.1 locus, typically not observed in the EBV-positive subtype. Our results support a role for EBV viral machinery in promoting tumorigenesis in the absence of the large-scale somatic lesions observed in EBV-negative cHL. These findings indicate the need to take into account EBV infection status when considering targeted therapies for the treatment of cHL, as many of the most common and targetable genetic lesions (including those in the JAK-STAT pathway) are seen nearly exclusively in the EBV-negative subtype and not the EBV-positive. Finally, the observation that germline homozygosity in *HLA-I* is enriched in EBV-positive cHL highlights a new risk factor for this disease. Future work should determine how germline *HLA-I* zygosity can predispose to the development of EBV-positive cHL and whether this may be a risk factor for other virus-associated cancers.

3.4 Methods

EBV infection status calling

EBV infection status of newly sequenced cHL patients was determined by standard EBER in-situ hybridization on fixed tissue sections and confirmed by presence of reads aligned to the EBV reference assessed by samtools idxstats.

Sample preparation and sequencing

Whole-genome sequencing (WGS) was performed on tumor and normal samples from a cohort of 27 cHL cases previously subjected to whole-exome sequencing (WES) [93], as well as

from 5 newly microdissected EBV-positive cHL cases. For each patient, we laser-microdissected HRS cells (n=1200-1800 per case) along with a similar number of adjacent non-neoplastic cells from frozen lymph node sections and subjected the samples to whole genome amplification (WGA) in duplicate [93]. Here, whole genome sequencing (WGS) was performed separately for each tumor duplicate at a median depth of 44X, as well as for the pooled normal duplicates to a median depth of 44X. For 5 cases subjected to WES, we additionally sequenced unamplified DNA from peripheral blood at a median depth of 41X. Preparation of libraries for WGS was done using Illumina TruSeq DNA PCR-Free library kit and Illumina TruSeq Nano DNA library kit for 31/32 cases and 1/32 case respectively, followed by paired-end sequencing for 2x125 cycles on Illumina HiSeq 2500 and for 2x150 cycles on Illumina NovaSeq instruments for 20/32 cases and 12/32 cases, respectively.

Single nucleotide and indel variant calling pipeline

Whole genome sequencing samples were aligned to GRCh37 using the Burrows-Wheeler aligner. Samples were pre-processed by indel realignment, duplicate removal, and base recalibration with GATK [120] following the GATK best practices workflow [121]. SAVI-v2 [122], an in-house variant caller, was used to call somatic variants. As there were two microdissected tumor and one normal sample for each case sequenced by WGS, we adapted the WES pipeline for one normal microdissected sample rather than two [93], and defined somatic variants present in a major tumor clone as those having a VAF >= 20% in both tumor replicates, <3% in the pooled normal microdissected sample, and <1% in the unamplified blood sample (when available) at any genome bases that were covered by at least 6 reads in all tumor, normal and blood samples of each case (representing a median of 89% of all genome bases, IQR 87-90%). The threshold of 6 reads was selected because, in a comparison to deeper WES data from the same

cases (n=31 with at least 1 mutation called on both WES and WGS) taken as benchmark (median coverage 148X), this threshold provided absolute specificity (99.999%) while preventing the loss of sensitivity at minimum depths higher than 6 (**Figure 3.S.6**). All other filters, including those to remove SNPs, strand bias, and homopolymeric indels, were applied as previously described. [93] To further account for errors in the whole genome sequencing data, we removed variants within ENCODE or Duke consensus blacklist regions (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability), as those are suspected artifacts. Finally, we used normal samples from all 32 WGS cases to construct a cohort supernormal and removed any putative somatic variants in these cases that was called also in the supernormal.

WES data were analyzed for 38 cHL cases (34 already published [93] plus 4 newly microdissected and processed as in ref. [93], except for using the updated Agilent SureSelectXT V6+Cosmic probes and kit), following a bioinformatics pipeline previously described [93] that was also applied to the WES data of 18 cHL cases available for download from ref. [107]. The latter data were generated after flow cytometry sorting of tumor and normal cells from cryopreserved tissue cell suspensions, without subsequent WGA, and comprised one tumor and one normal sample per case. These 56 WES samples were subjected to the same bioinformatics pipelines for mutation calling (described in ref. [93]) and for copy number aberration calling and *HLA-I* analyses (described further below).

Mutation signature analysis

Mutation signatures were called from clonal somatic variants using Palimpsest [123], an NMF-based mutation signature caller. Unsupervised mutation signature analysis was first run for combined single base substitution (SBS), double base substitution (DBS), and insertion-deletion (ID) signature calling. To improve detection of potential AID-associated signatures expected in

this data set due to the germinal center B cell of origin of this tumor, *de novo* signatures were also separately called for clustered variants, i.e. variants grouped by nearest mutation distance (NMD) below a threshold defined according to approach of ref. [81], i.e. NMD < 316 bases for WGS based on visual inspection of the histogram of NMD (**Figure 3.S.5**). All *de novo* signatures found in cHL cases were compared to published whole genome amplification signatures [124] to rule out sequencing artifacts from the amplification procedure. Supervised mutation signature analysis was then run on using COSMIC v3 mutation signatures that were the highest ranking match for the inferred signature from the *de novo* analysis, based on cosine similarity score.

Mutation hot spots

Mutation hot spots in cHL genomes were identified using Fishhook [112], an R package that uses Gamma-Poisson regression to determine genomic intervals enriched or depleted for somatic mutations. The hg19 genome was binned into 10,000 base pair windows. Regions of low mappability, low complexity, and sites of unusually high numbers of aberrant SNP calls from 1,000 Genomes Project were masked using the "universal mask" for whole genome sequencing variant calling [125] as described by Imielinski et al. [112] GC-fraction was included as a covariate. Hot spots were called separately for single nucleotide variants and indel variants.

Aberrant somatic hypermutation (ASHM) analysis

ASHM-associated regions were identified as regions within 2 kb of the transcription start site of 126 previously identified targets of ASHM [75, 126, 127], as previously described [33]. Mutations were included in the analysis if they passed the single nucleotide and indel variant calling pipeline described above. "ASH mutations" were defined as somatic point mutations occurring within ASHM-associated regions. "Mutated ASHM genes" were counted as the number of unique Ensembl gene IDs containing at least one ASH mutation.

Copy number segmentation and variant calling

Copy number segmentation of cHL samples was conducted using Control-FREEC [128] for each pair(s) of tumor and normal samples sequenced by whole genome sequencing available for each case. For patients with more than one tumor and/or normal sample, only aberrations occurring in overlapping regions of gain or loss present in both tumor replicates were counted as non-normal copy number. Copy number alterations (CNAs) were defined based on the following absolute CN cut off values: CN > 2.3 gain, CN > 3.6 amplification, CN < 1.7 heterozygous loss, and CN < 0.8 homozygous loss. The CN of overlapping regions of CNA was calculated as the mean of the constituent copy numbers. When compared to fluorescence in situ hybridization data already available for JAK2 and TNFAIP3 deletion in 33/34 and 32/34 previously published cases, sensitivity and specificity of this analysis for focal JAK2 gain and TNFAIP3 loss on WGS cases were 38% and 94% for JAK2 copy number gain, and 29% and 81% for TNFAIP3 deletion. Structural variants of Hodgkin WGS caes were called using manta [129] and high-confidence variants were selected through manual review.

In order to define significant regions of recurrent CNAs in Hodgkin lymphoma, GISTIC 2.0 [130] was applied to the copy number segmentation of 56 cases of cHL sequenced by WES with a $q < 0.2$, a maximum segmentation threshold of 10,000, and all other parameters default via the GenePattern server (https://www.genepattern.org/). To denoise the list of putative recurrent CNAs in the WES cases, GISTIC 2.0 was applied to the cohort of 32 cases of cHL sequenced by WGS, with a $q < 0.5$, a maximum segmentation threshold of 10,000, and all other parameters default via the GenePattern server. The list of putative CNA regions defined by WES was filtered to include only those peaks occurring on the same cytoband as a peak called in the WGS cohort, and/or in a region previously described as a known site of recurrent copy number aberration in B

cell lymphomas by manual curation. Significant peaks were defined as those having a q* < 1e-04, where q* = sqrt(q_{WES} * q_{WGS}), where q_{WES} is the q value of the peak in the cytoband called by GISTIC in the WES cohort and q_{WGS} is the q value of the peak in the cytoband called by GISTIC in the WGS cohort. GISTIC peaks of gain or amplification were counted as present in a patient if the maximum inferred tumor copy number within the wide peak limit region was > 2.3 (gain) or > 3.6 (amplification). GISTIC peaks of deletion were counted as present in a patient if the minimum inferred tumor copy number within the wide peak limit region was < 1.7 (heterozygous loss) or < 0.8 (homozygous loss). Arm and whole-chromosome level CNAs were defined as lesions of the same type (i.e. gain or loss) that covered >75% of the chromosome arm or chromosome, respectively.

Class I *HLA* allele typing and mutation calling

Class I *HLA* allele typing was performed using PolySolver [104] with default parameters. When available, alleles were called from sequenced unamplified blood samples. Otherwise, cases were called as homozygous only if they were called as homozygous by PolySolver in all tumor and normal samples. “Homozygous” refers to cases where both inferred alleles of an *HLA-I* gene are the same to the two-field resolution (allele group and specific HLA protein). *HLA* mutation calling was performed using PolySolver. *HLA* loss of heterozygosity in non-whole genome amplified cases was assessed with LOHHLA with a minimum coverage threshold of 5 reads using purity and ploidy inferred by Sequenza [131] with default parameters. Cases were called as having *HLA* loss of heterozygosity if the p value for loss of the allele was < 0.01 and the inferred copy number of the allele was < 0.5 (in both tumor samples, when applicable), following thresholds used in McGranahan et al [132]. *HLA* loss of heterozygosity was evaluated only for patients that

did not undergo whole genome amplification (n=18) because allele dropout from the amplification procedure prevented accurate *HLA-I* loci allelic copy number calling in those samples.

Statistical Analyses

Analyses of significance of mutation counts and frequencies were performed using the Mann-Whitney U test and Fisher's exact test, respectively. Significance of mutation signature activities were assessed using Student's t test. Multiple hypothesis correction was applied using the Benjamini-Hochberg Procedure.

3.5 Supplementary Figures

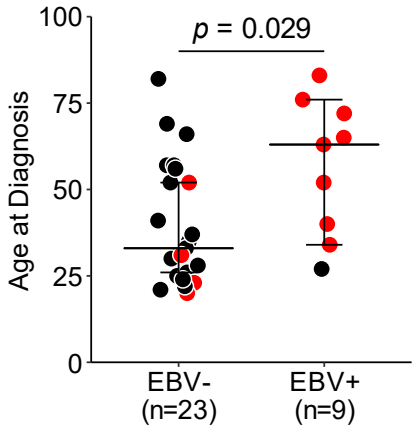


Figure 3.S.1. Age at diagnosis of 32 cHL sequenced by WGS. The age of the EBV-positive individuals was significantly higher than the EBV-negative (p=0.029, MWU test). Mixed cellularity cases are colored in red.

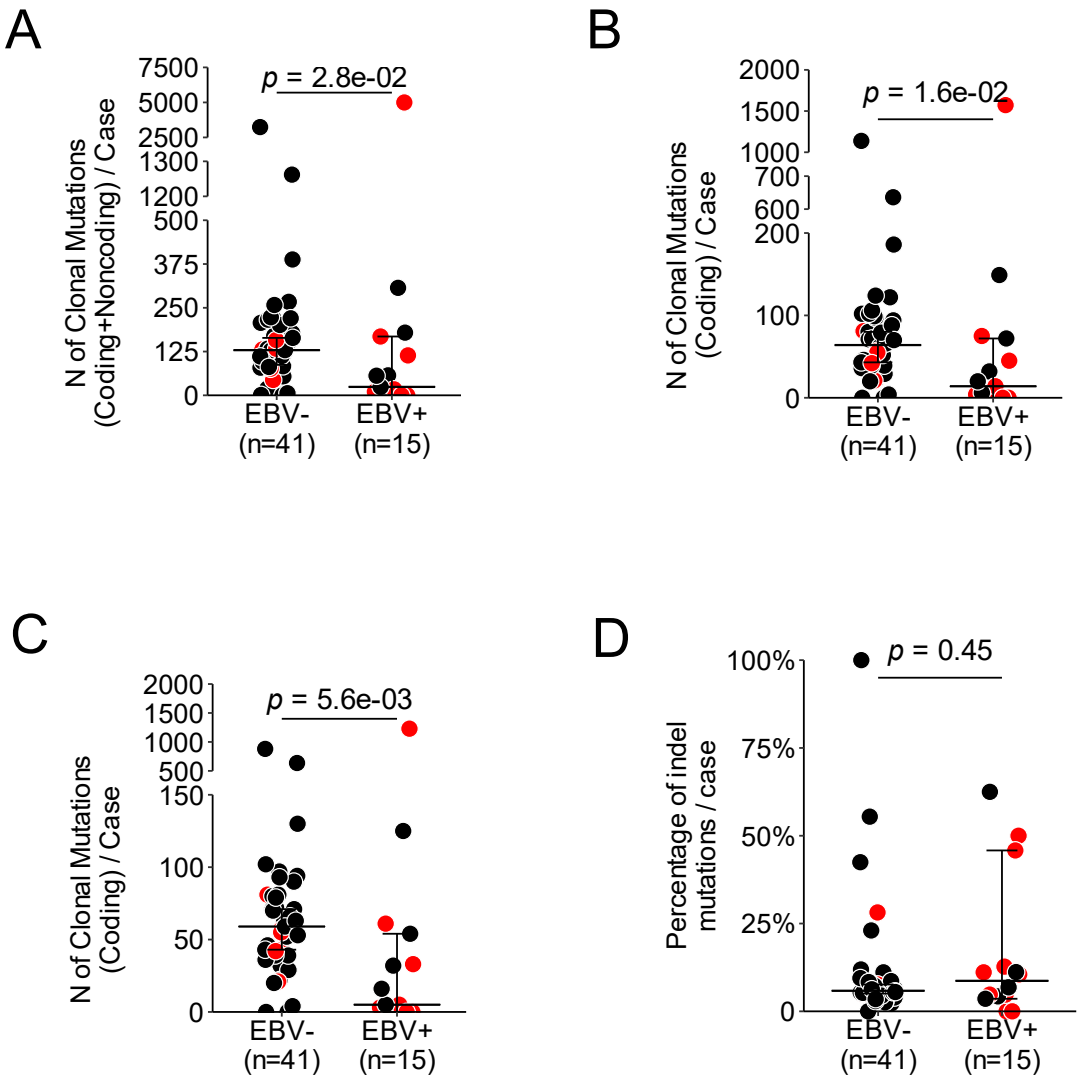


Figure 3.S.2. EBV-negative cHL exomes harbor a greater number of somatic mutations compared to EBV-positive cHL. The mutation load for each sample sequenced by WES was calculated as A) the total number of clonal mutations (point mutations and short indels), B) the total number of clonal mutations in coding regions, and C) the total number of clonal, nonsilent mutations in coding regions. D) Proportion of clonal mutations that are insertions/deletions in each case. In all panels, p-values were computed with the Mann-Whitney U test, and mixed cellularity cases are colored in red.

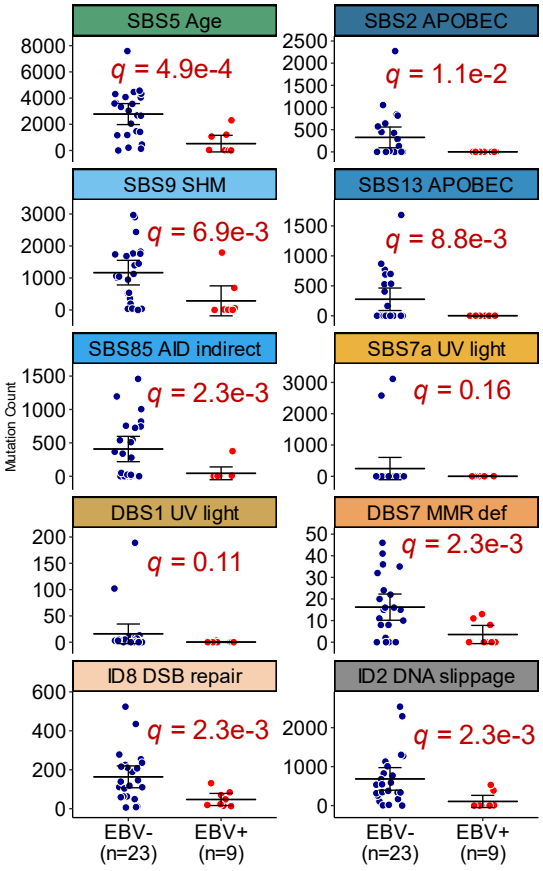


Figure 3.S.3. Counts of mutations attributed to each mutation signature identified from WGS of 32 cHL. q values from Student's t test, BH corrected.

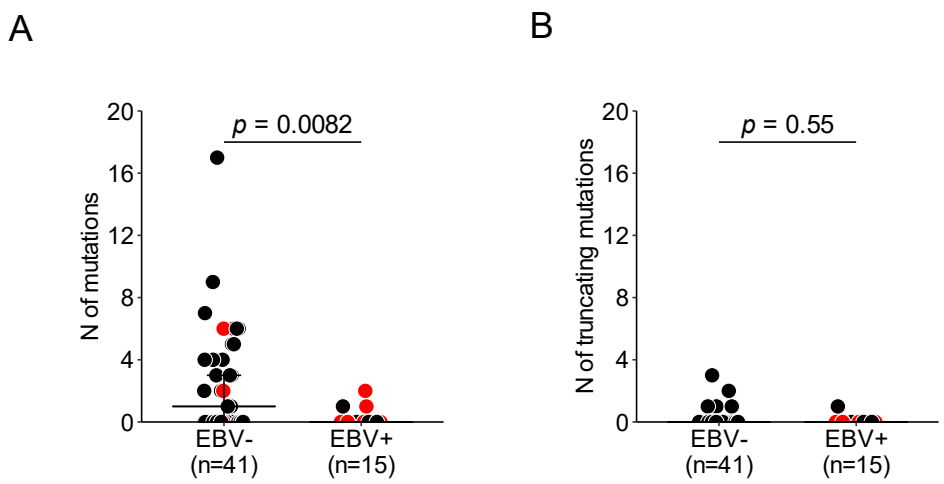


Figure 3.S.4. Somatic mutations in *HLA-I*. A) Count of total somatic mutations in *HLA-I*. B) Count of truncating somatic mutations in *HLA-I*. p values: MWU test. Mixed cellularity cases are colored in red.

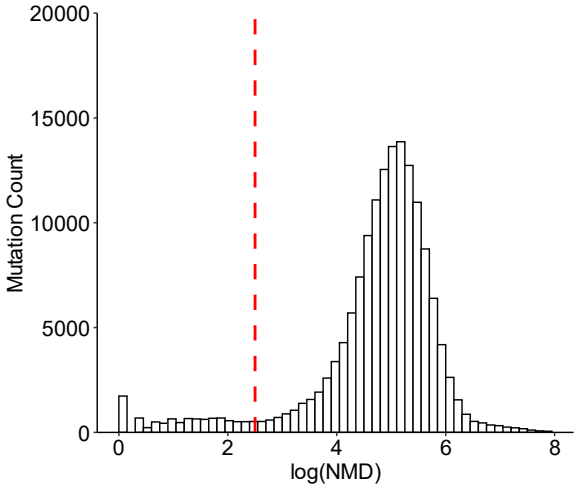


Figure 3.S.5. Histogram of mutation count by nearest mutation distance (NMD) in 32 cHL. Approximate point between two peaks of NMD (red, dashed line) was used to determine cut-off for cHL de novo mutation signature analysis.

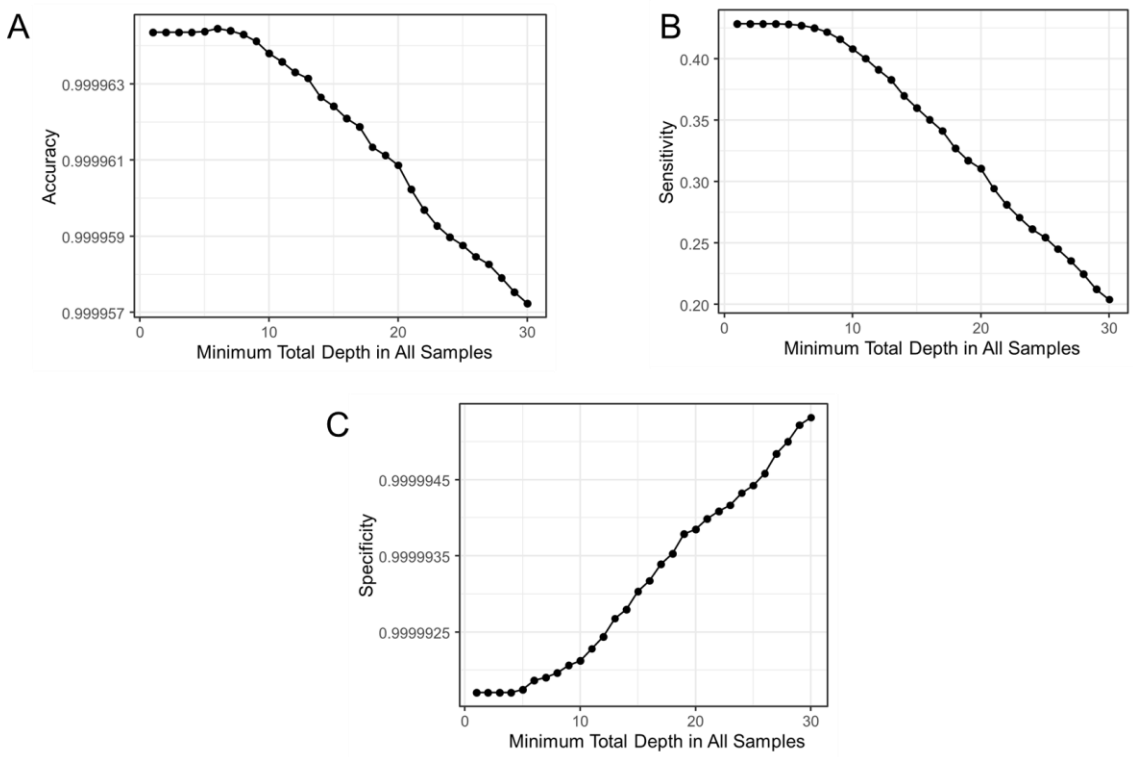


Figure 3.S.6. Accuracy of WGS mutation calls assessed with paired WES samples from the same patient. A cut-off value of a minimum of 6 total reads in all samples was chosen for somatic mutation calling in WGS samples as this maximized accuracy (A) while minimizing loss of sensitivity (B). Specificity also shown (C).

Chapter 4: Genomic landscape of virus-associated cancers[3]

4.1 Introduction

An estimated 15-20% of cancers are attributable to infections [42, 43], and 8-10% are caused by viruses [44, 133]. To date, seven viruses are known to be associated with the development of cancers in humans (oncoviruses): human gammaherpesvirus 4 (HHV-4, also known as Epstein-Barr virus [EBV]), human herpesvirus 8 (HHV-8), human papillomavirus (HPV), human T-cell lymphotrophic virus type 1 (HTLV-1), hepatitis B virus (HBV), hepatitis C virus (HCV), and Merkel cell polyomavirus (MCPyV) [134]. The first human oncovirus to be described was EBV, following the discovery of viral particles in cultured lymphoblasts from Burkitt lymphoma (BL) in 1964 [135]. Since then, EBV has been linked to a wide array of both hematological and solid tumors, including Hodgkin lymphoma (HL) (20-50% of cases [45]), other B and T cell lymphomas (such as plasmablastic lymphoma [PBL], extranodal NK T-cell lymphomas [NKTCL], and primary central nervous system lymphoma [PCNSL] in immune-suppressed patients, each in 70-100% of cases [56-58]), gastric cancer (GC) (8.7% of cases [59]), and nasopharyngeal carcinoma (NPC) (100% of cases [60]), summing to approximately 1% of all cancers [44]. It has also been suggested that EBV may be associated to B-cell lymphomagenesis more widely than currently acknowledged, possibly via a "hit and run" mechanism [136]. By the 1980s, HPV, HBV, HCV, and HTLV-1 had been identified as additional oncoviruses. High-risk HPV types such as HPV16 and HPV18 contribute to the pathogenesis of cervical cancer (CC) (95% of cases [51]), head and neck squamous cell carcinoma (HNSCC) (30% [52]), and anogenital

[3] Material in this chapter is contained wholly or in part in a manuscript in preparation by Karen Gomez*, Gianluca Schiavoni*, Yoonhee Nam, Jean-Baptiste Reynier, Cole Khamnei, Michael Aitken, Giuseppe Palmieri, Antonio Cossu, Arnold Levine, Carel van Noesel, Brunangelo Falini, Laura Pasqualucci, Enrico Tiacci**, and Raul Rabadan**
*Contributed equally. **Jointly supervised this work.

cancers (70-90% [137]), representing 5% of all cancers [137]. HBV and HCV have been associated with up to 56% and 20% of hepatocellular carcinomas (HCC) [53] and 2% and 1% of total cancers [44], respectively. HTLV-1, the only oncogenic retrovirus yet described, is necessary for the development of adult T-cell leukemia/lymphoma [62]. Since the discovery of HHV-8 in AIDS patients with Kaposi sarcoma (KS) in 1994 [61], HHV-8 has been implicated in the pathogenesis of Kaposi sarcoma (100% of cases), primary effusion lymphoma (100% [138]), and Castleman's disease (20-40% [139]). In 2008, MCPyV was linked to Merkel cell carcinoma (MCC) [140], and has since been identified as an etiological agent in 80% of MCC tumors [63]. Recently, HPV42, previously classified as a "low-risk" HPV type, has been found in a majority of digital papillary adenocarcinomas [141]. It is suspected that viruses may play a causative role in the pathogenesis of other cancer types [142, 143], and there may be more oncoviruses that have yet to be discovered.

While the mechanisms of malignant transformation caused by oncoviruses differ, there are some general patterns that are observed [144]. First, oncoviruses cause a persistent, long-term infection, and tumors develop years after the initial infection. For example, most individuals are infected with EBV by early childhood (in developing countries) or adolescence (in developed countries) [145], but an EBV-associated cancer may not develop until old age. Hepatocellular carcinoma develops 10-30 years after infection with HBV or HCV [146], and cervical cancer develops 25-30 years following infection with HPV [147]. Second, oncoviruses encode proteins that directly contribute to malignant transformation. In HPV infected cells, the E6 and E7 oncoproteins inhibit the tumor suppressors p53 and Rb, respectively [148]. The vGPCR protein encoded by HHV8 induces angiogenesis and promotes cell transformation [149]. EBV expresses different genes depending on the viral latency program. In Burkitt lymphoma, EBV expresses a type 1 latency program including the protein EBNA-1, which is necessary for the replication of

viral DNA and may inhibit apoptosis [150]. In Hodgkin lymphoma, gastric carcinoma, and nasopharyngeal carcinoma, EBV expresses a type II latency program, including EBNA-1 as well as the proteins LMP1 and LMP2, which activate the NF-κB and PI3K-AKT pathways [150]. Third, viral infection is necessary, but not sufficient for malignant transformation. Many oncoviruses are highly prevalent in the general population: 90-95% of people worldwide are infected with EBV [145], 80% of individuals will acquire an HPV infection by age 45 [151], and MCPyV is detected in 80% of individuals in the general population by age 50 [152]. However, only a small fraction of those infected with oncoviruses will develop cancer, suggesting additional genetic and/or environmental factors are required.

The factors that contribute to the malignant transformation of virus-infected cells remain incompletely understood, but are known to include a combination of environmental, immune, inherited, and somatic components. While many of these components have been described for individual cancer types, relatively little has been reported about the clinical and genetic factors that are common across virus-associated cancers. In this study, we investigate virus-associated oncogenesis through an integrative analysis of nine cancers for which a subset of cases are associated with five oncoviruses through data from both published and newly collected data sets. We identify patterns of common phenotypic characteristics, somatic drivers, germline risk factors, and therapeutic responses among these malignancies. This study provides a comprehensive analysis of human cancers that develop in the context of viral infection and key factors related to their pathogenesis.

4.2 Results

Virus-associated cancer show unique epidemiological trends

Virus-associated cancers are known to follow unique epidemiological patterns compared

to non-virus-associated cancers. For example, the age of incidence of HL follows a bimodal distribution which reflects two distinct histological subtypes with different etiologies: 1) nodular sclerosis, usually EBV-negative, in young adults, and 2) mixed cellularity, often EBV-positive, in older adults [153, 154]. In order to illustrate other common demographic characteristics of virus-associated malignancies, we analyzed data from the Global Cancer Observatory (GLOBOCAN 2020) [155] and published incidence rates in 82 studies of 11 cancer types linked to 7 viruses and 13 non-virus-associated cancers [56, 156-236]. First, we compared the incidence rates of viral cancers in males versus females (M/F) reported in select published studies [56, 156-236] (**Figure 4.1A**). We found that the M/F ratio reported was greater overall in virus-associated cancers compared to nonviral cancers ($p=0.03$, Mann Whitney U [MWU] test). Among studies that reported rates of male and female incidence for virus-positive and virus-negative tumors specifically, virus-positive cases tended to have a greater M/F ratio than virus-negative cases ($p=2.4e-10$). This trend was also consistent when separately comparing virus-positive and virus-negative tumors of gastric cancer ($p=1.04e-10$) and Hodgkin lymphoma ($p=0.015$). One exception to this trend is MCPyV-positive MCC, which has a lower M/F ratio compared to virus-negative MCC. Interestingly, digital papillary adenocarcinoma, which has been recently associated to HPV42, is more frequent in males compared to females at a ratio of 4 to 1 [237].

To examine how the incidence of virus-associated cancers differs by geographic location, we compared the age-standardized incidence rates (ASR) of 4 cancers in 185 countries reported in GLOBOCAN 2020. To identify countries with high ASR of virus-positive tumors for different virus-associated cancers, we estimated the number of cases attributable to viral infection by country from GLOBOCAN 2020 total incidence counts and the attributable fraction per region estimated by de Martel et al [44]. In HL, EBV-positive cases occur most frequently in North

Africa, the Middle East, and South America, with the lowest incidence occurring in East Asia

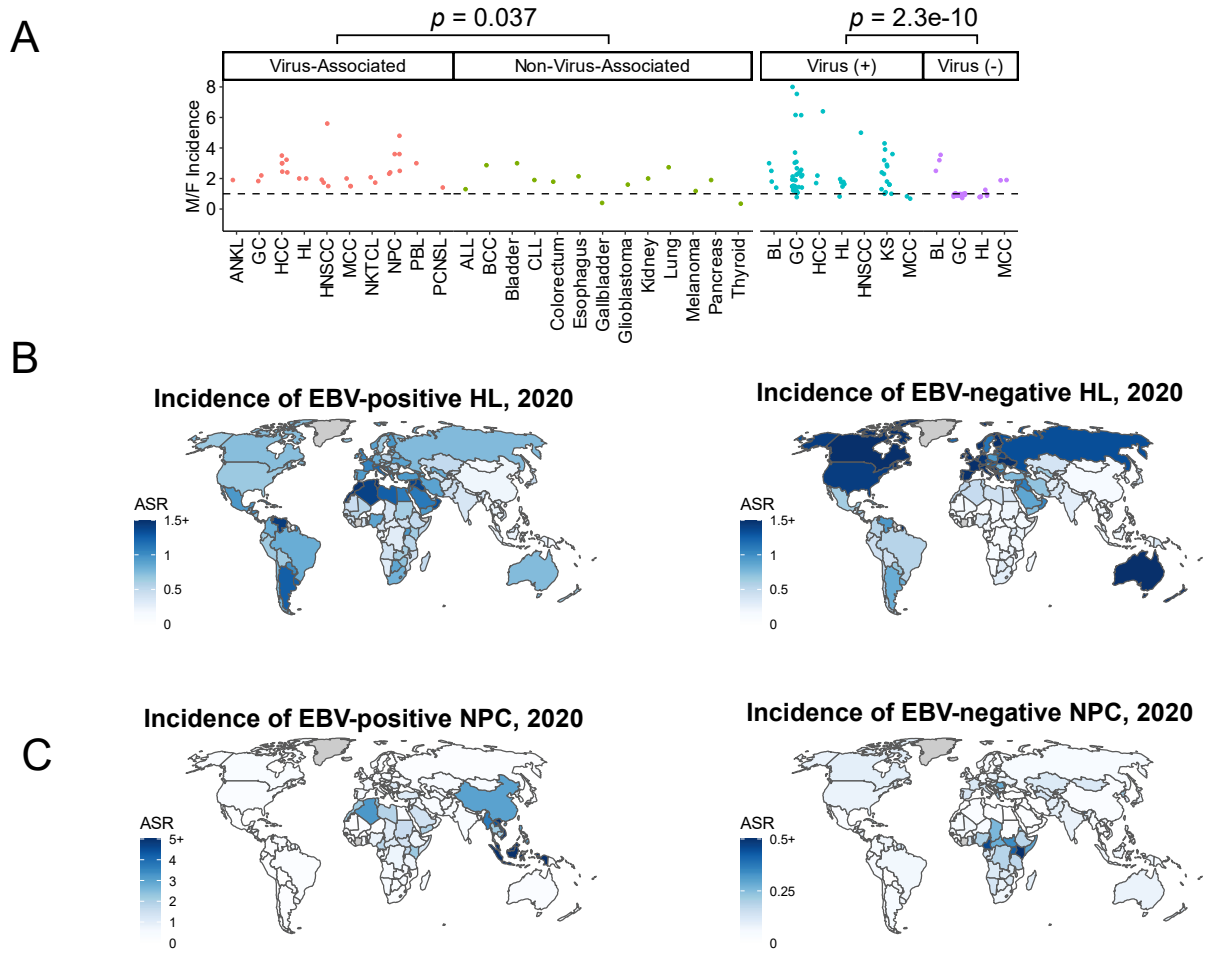


Figure 4.1: Epidemiological trends of virus-associated cancers. A) Incidence rates of virus-associated and non-virus-associated cancers (left) and virus-positive and virus-negative tumors in virus-associated cancers (right) in males compared to females (M/F) reported in selected published studies. Each point corresponds to an incidence ratio reported in a published study. Virus-associated cancers: ANKL, aggressive NK-cell leukemia (n=1); GC, gastric cancer (n=2); HCC, hepatocellular carcinoma (n=6), HL, Hodgkin lymphoma (n=2); HNSCC, head and neck squamous cell carcinoma (n=4), MCC, Merkel cell carcinoma (n=3); NKTCL, Natural killer/T-cell lymphoma (n=2); NPC, nasopharyngeal carcinoma (n=6); PBL, plasmablastic lymphoma (n=1); PCNSL, primary central nervous system lymphoma (n=1). Non-virus-associated cancers (n=1 each): ALL, acute lymphoblastic leukemia; BCC, basal cell carcinoma; CLL, chronic lymphocytic leukemia. Virus-positive tumors: BL, Burkitt lymphoma (n=4); GC (n=33); HCC (n=3); HL (n=6); HNSCC (n=1); KS, Kaposi sarcoma (n=15); MCC (n=2). Virus-negative tumors: BL (n=3); GC (n=31); HL (n=6); MCC (n=2). B) Estimated incidence rates of EBV-positive HL (left) and EBV-negative HL (right) by country. C) Estimated incidence rates of EBV-positive NPC (left) and EBV-negative NPC (right) by country.

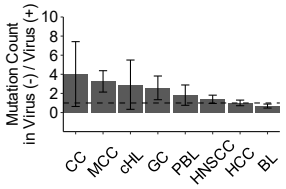
(**Figure 4.1B**). In contrast, most cases of EBV-positive NPC occur in China and southeast Asia (**Figure 4.1C**). Similarly, Kaposi sarcoma and cervical cancer (nearly all of which are virus-positive) show disparities in incidence by geographic location (**Figure 4.S.1**). These results illustrate that the locations of global hot spots of virus-positive tumor incidence vary by virus and even among cancers associated with the same virus. These disparities reflect differences in risk factors for virus-positive tumors among human populations, both genetic (e.g. inherited susceptibility polymorphisms [238, 239]) and environmental (e.g. oncovirus prevalence [240], and lifestyle factors such as smoking or diet, which affect overall cancer risk [241]).

Virus-positive tumors have fewer somatic mutations than virus-negative tumors

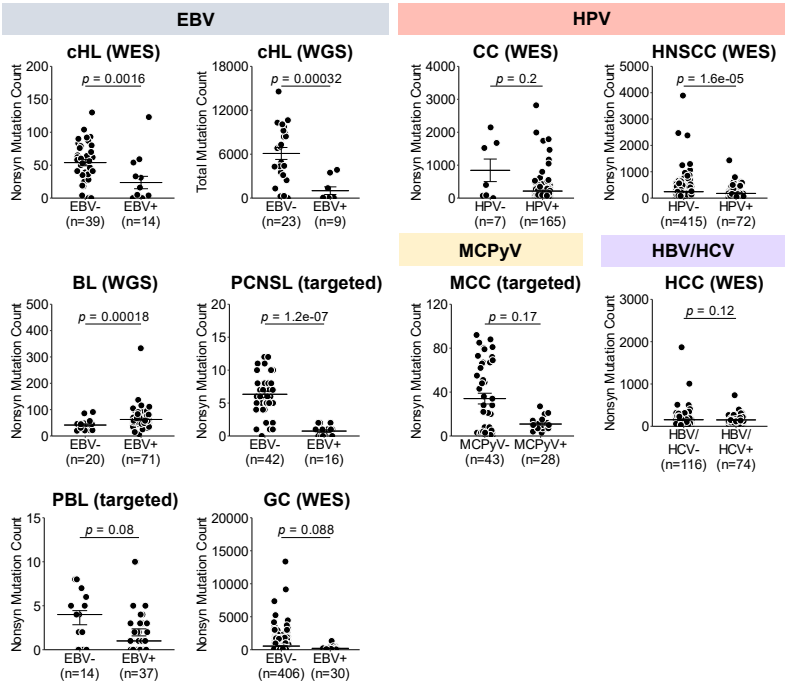
In order to quantify the somatic mutation burden of virus-associated cancers, we aggregated somatic mutation data from 1,658 tumors reported in published studies of nine cancers sequenced by whole exome sequencing (classical HL [cHL] [93, 107] [n=56], PBL [101] [n=15], GC [242] [n=440], HCC [243] [n=196], CC [244] [n=178], and HNSCC [245] [n=487]), targeted DNA sequencing (PBL [101, 246] [n=36], PCNSL [247] [n=58], MCC [248] [n=71], and BL [249] [n=29]), and/or whole genome sequencing (BL [249] [n=91] and newly sequenced cHL [n=32] [see Methods]). In general, virus-negative tumors had a higher count of nonsynonymous mutations compared to virus-positive tumors (**Figure 4.2A**). The count of nonsynonymous mutations was significantly lower in virus-positive compared to virus-negative cHL sequenced by WES (median 4.5 compared to 56, respectively, p=0.0016, MWU test), PCNSL (median 1 and 6, p=1.2e-7), and HNSCC (median 94.5 and 168, p=1.6e-5), and trended towards significance in PBL (median 1 and 4, p=0.080), GC (median 131.5 and 169, p=0.088), and to a lesser extent in CC (median 105 and 399, p=0.20), and MCC (median 10 and 28, p=0.17) (**Figure 4.2B**). While the mutation load in BL was higher in virus-positive tumors (median 53 and 42, p=0.00018), the count of

nonsynonymous mutations in genes previously described as BL drivers [249] trended towards lower in the virus-negative (median 8 and 6, p=0.074) (**Figure 4.S.2**). Furthermore, when restricting the analysis to driver genes, the higher mutation count in virus-negative versus virus-positive cases became statistically significant in MCC (median 5 and 1, p=0.029) (**Figure 4.S.2**). Overall, the total mutation count and/or driver mutation count was lower in virus-positive compared to virus-negative tumors in 8/9 cancers studied (**Figure 2C**). The exception to this trend was HCC, which had a higher total nonsynonymous mutation count (median 129 vs 118, p=0.12) and driver mutation count (median 2 vs 1, p=0.0017) in virus-positive versus negative cases.

A



B



C

Virus	Cancer	count of nonsyn mutations in virus (+)	count of nonsyn mutations in driver genes in virus (+)
EBV	cHL	lower	lower
	BL	higher	lower
	PBL	lower	lower
	PCNSL	lower	NA
	GC	lower	higher
HPV	CC	lower	lower
	HNSCC	lower	lower
MCPyV	MCC	lower	lower
HBV/ HCV	HCC	ND	higher

Figure 4.2: Mutation burden of virus-positive and virus-negative tumors in nine cancers. A) Ratio of average number of somatic nonsynonymous mutations in virus-negative tumors compared to virus-positive tumors. MCC (n=71), HNSCC (n=487), CC (n=172), cHL (n=54), GC (n=436), BL (n=91), PBL (n=51), HCC (n=190), and PCNSL (n=58). B) Counts of somatic nonsynonymous mutations in virus-positive and virus-negative tumors in the same cancers. C) Summary of trends in mutation load in virus-positive versus virus-negative tumors by cancer type. Results highlighted in orange and dark blue reach significance past a threshold p <0.05, while results highlighted in light blue indicate a trend that does not reach significance p<0.05 (MWU test).

Virus-associated cancers display unique mutation signatures

To detect and quantify the relative contribution of COSMIC mutation signatures [250] within the nine virus-associated cancers, we next applied a supervised non-negative matrix factorization approach informed by *de novo* signature calling. Virus-positive tumors exhibited different activities of mutation signatures compared to virus-negative tumors of the same cancer type (**Figure 4.3, Figure 4.S.3**). In cHL, the detected mutation signatures included SBS5/age, SBS9/somatic hypermutation (SHM), SBS85/activation-induced cytidine deaminase (AID), SBS2/13/apolipoprotein B mRNA editing enzyme, catalytic polypeptide (APOBEC) (**Figure 4.3A, left**). The absolute count of mutations ascribed to each of these signatures was significantly lower in EBV-positive versus EBV-negative cases (**Figure 4.S.3A**). For SBS2/APOBEC, the relative signature activity (i.e., the proportion of signature mutations among all mutations observed in each case) was lower in EBV-positive compared to EBV-negative tumors (mean proportion 0.0079 and 0.047, respectively, q=0.11) (**Figure 4.3A, right**). APOBEC enzymes belong to a family of cytidine deaminases that includes AID, and a previous study focused on a few selected genomic regions showed that the AID-mediated process of SHM functions aberrantly in cHL [111]. Our unbiased analysis revealed that the inferred SHM signatures SBS9 and SBS85 had higher absolute activity in EBV-negative versus EBV-positive cases genome-wide (**Figure 4.S.3A**). Additionally, we found that mutations occurring in regions within 2 kb of the transcription start site of 126 genes known to be targeted by aberrant SHM (ASHM) in diffuse large B-cell lymphomas (DLBCL) [75, 126, 127] were enriched in EBV-negative compared to EBV-positive cHL (median 1 and 0, p=0.045, MWU test; **Figure 4.3A, right**). Collectively, these data suggest there is a greater selective pressure for somatic mutations induced by APOBEC and AID activities in EBV-negative cHL, that in EBV-positive cHL might be substituted by oncogenic activities of

viral proteins.

UV light is known to be the major etiological agent of MCC tumors in the absence of viral infection [248]. Accordingly, the absolute count of mutations attributed to SBS7a/7b/UV light was lower in MCPyV-positive compared to MCPyV-negative (**Figure 4.S.3B**). MCPyV-positive MCC cases also displayed a lower proportion of mutations associated with each UV light signature (SBS7a/b) compared to MCPyV-negative cases (mean 0.044 and 0.13, q=0.023 and mean 0.11 and 0.28, q=0.0019, respectively) (**Figure 4.3B**).

In GC, among mutation signatures differentially active by virus status (**Figure 4.3C** and **Figure 4.S.3C**), of potential interest is the greater relative and/or absolute activity of SBS20, SBS15, and SBS21 mismatch repair deficiency signatures in EBV-negative versus EBV-positive cases. For example, mean relative activity of SBS15 is greater in EBV-negative than EBV-positive cases (mean 0.15 and 0.076, respectively; q=0.00026). These findings suggest a greater role for microsatellite instability (MSI) in the pathogenesis of EBV-negative GC. MSI as assessed by standard methods is a defining characteristic of a GC subtype that is exclusively EBV-negative and comprised 73/406 (18%) of EBV-negative patients in the TCGA cohort. Accordingly, the relative activities of SBS20, SBS15, and SBS21 were greater in the conventionally defined MSI subtype compared to the other EBV-negative cases (mean 0.11 and 0.010, p<2.2e-16; mean 0.30 and 0.12, p<2.2e-16; and mean 0.071 and 0.0071, p=4.86e-9, respectively). The relative activity of SBS15 was also greater in non-MSI EBV-negative compared to EBV-positive tumors (mean 0.12 and 0.076, p=0.021), suggesting that MSI may be more widespread in EBV-negative GC than currently appreciated with standard methods.

There was no significant difference in the absolute counts or proportions of signatures in HPV-positive versus -negative CC, likely due to the limited number (n=7) of HPV-negative cases

reported in TCGA (**Figure 4.3D**). However, in HNSCC, HPV-negative cases had a greater number of ID3 and DBS2 mutations related to smoking (**Figure 4.S.3D**), a known risk factor for HNSCC that may be less relevant for HPV-driven carcinogenesis [251]. In contrast, HPV-positive HNSCC had higher absolute and relative activity of SBS2/APOBEC (**Figure 4.S.3D and Figure 4.3E**), a finding consistent with the hypothesis that HPV oncoproteins may increase APOBEC3A and APOBEC3B expression and mutagenic activity [251, 252].

In HCC, there was no difference in absolute count of mutations attributed to mutation signatures, consistent with the similar mutation burden in virus-positive and -negative HCC overall. However, HCC tumors positive for HBV and/or HCV had a greater proportion of mutations due to SBS24/aflatoxin, an environmental carcinogen known to predispose to HBV/HCV-mediated cirrhosis [253, 254] (mean 0.16 and 0.085 q=0.021) (**Figure 4.3F**).

In BL, we found there was a greater count of mutations in exonic regions attributed to SBS17b/unknown etiology in EBV-positive compared to EBV-negative cases (**Figure 4.S.3E**), consistent with a previous study of genome-wide mutation signatures in this cohort [249]. Similarly, EBV-positive BL had a greater proportion of mutations due to SBS17b/unknown etiology compared to EBV-negative BL (mean 0.031 and 0.0030, q=0.0015) (**Figure 4.3G**). We additionally found that signatures SBS6/MMR and SBS5/age contributed to a greater count of mutations in exonic regions in EBV-positive compared to EBV-negative BL (**Figure 4.S.3E**). Overall, these results illustrate that the signatures of somatic mutation processes vary depending on infection status for each cancer, highlighting differing selective pressures on the cancer genomes in the presence or absence of viral oncoproteins.

Virus-associated tumors harbor frequent mutations in RNA helicases *DDX3X* and *EIF4A1*

In order to identify genomic loci that are preferentially mutated in virus-positive tumors,

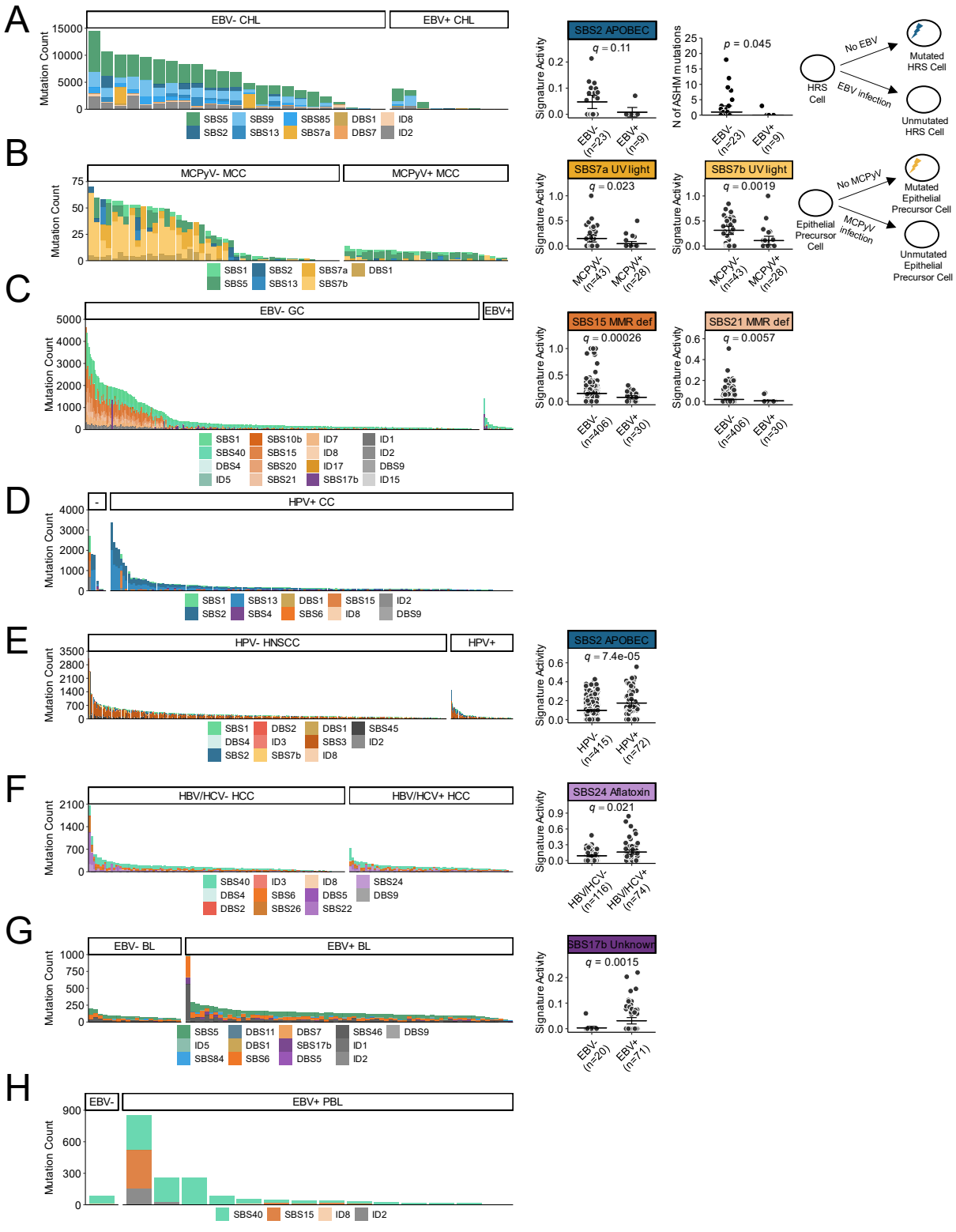


Figure 4.3: Mutations signatures in eight virus-associated cancers. A) cHL (n=32), B) MCC (n=71), C) GC (n=436), D) CC (n=172), E) HHC (n=190), F) HNSCC (n=487), G) BL (n=91), H) PBL (n=23). Activities of signatures in virus-positive compared to virus-negative cases

are shown for signatures with a significant difference in activity (q<0.05) and/or biologically relevant trend (right). Schematic representation of the effect of the absence of processes behind key mutation signatures in virus-associated cases is shown in A) cHL (EBV-positive) and B) MCC (MCPyV-positive).

we compared the rate of nonsynonymous mutations in the pooled cohort of 537 virus-positive tumors and 1,121 virus-negative tumors from 9 cancer types. We found that four genes had an elevated odds of mutation in virus-positive tumors compared to virus-negative tumors: *EIF4A1* (OR 69.43, 95% CI 4.15-1160.26, q=2.37e-5, MWU test, BH corrected), *DDX3X* (OR 7.07, 95% CI 4.06-12.31, q=2.25e-11), *ARID1A* (OR 2.49, 95% CI 1.72-3.58, q=7.1e-4), and *MYC* (OR 3.92, 95% CI 2.57-6.01, q=8.67e-8), although the latter was driven by GC exclusively (OR 6.13, 95% CI 0.54-70.01, q=0.38) (**Figure 4.4A**). When looking at individual cancer types, *EIF4A1* had a significant odds of mutation in EBV-positive GC compared to EBV-negative GC (OR 63.09, 95% CI 2.94-1352.57, q=4.21e-6) and showed a similar trend in BL (OR 7.92, 95% CI 0.45-140.54, q=0.12) (**Figure 4.S.4**). *DDX3X*, an RNA helicase in the same family as *EIF4A1*, had a nominally elevated OR of mutation in EBV-positive cHL (OR 8.78, 95% CI 0.34-228.59, q=0.34), HNSCC (OR 3.73, 95% CI 0.61-22.85, q=0.34), BL (OR 1.86, 95% CI 0.76-4.53, q=0.34), and GC (OR 1.64, 95% CI 0.083-32.72, q=0.74), though no cancer reached significance individually. *ARID1A* was significantly mutated in EBV-positive GC (OR 15.11, 95% CI 6.16-37.05, q=3.30e-12) but also trended towards an elevated OR in virus-positive HCC (OR 2.86, 95% CI 0.99-8.28, q=0.087), BL (OR 1.70, 95% CI 0.67-4.30, q=0.35), and cHL (OR 1.42, 95% CI 0.12-17.03, q=0.81). Conversely, *TP53* had an elevated odds of mutation in virus-negative tumors compared to virus-positive tumors (OR 8.58, 95% CI 6.40-11.50, q=5.8e-51) (**Figure 4.4A**), which was significant for most cancer types individually (**Figure 4.S.4)**. Analysis of recurrent copy number aberrations also revealed recurrent loss of 9p21.3 (*CDKN2A*, *CDKN1A*) in virus-negative tumors (**Figure 4.S.5**).

EIF4A1 and *DDX3X* encode RNA helicases of the DEAD (Asp-Glu-Ala-Asp) box protein family which are known to play a role in splicing, RNA export, and cap-dependent translation initiation [255]. In order to further explore the role of these genes in virus-associated cancers, we expanded the analysis of *EIF4A1* and *DDX3X* mutations to include 316 tumors from cancers that are virus-associated in almost 100% of cases: KS (10 newly sequenced cases), aggressive NK cell lymphoma (ANKL) [256] (n=14), ATL [257] (n=81), NKTCL [258] (n=100), and NPC [259] (n=111), for a total of 1,974 cases. In all, we identified 135 *DDX3X* nonsynonymous mutations (100 in virus-positive; 35 in virus-negative cases) and 27 *EIF4A1* nonsynonymous mutations (22 in virus-positive; 5 in virus-negative cases).

Among virus-positive tumor samples, somatic mutations in *DDX3X* were detected in 53% (50/93) BL, 29% (4/14) ANKL, 19% (19/100) NKTCL, 7% (1/14) cHL, 5% (4/81) ATL, 3% (1/30) PBL, 3% (2/60) HNSCC, and 2% (3/144) of CC. Mutations in *DDX3X* tended to occur in the helicase domain residues more frequently than expected by chance (p=4.0e-4, binomial test) suggesting selective pressure for a functional role (**Figure 4.4B**). This was similar for virus-positive cases only (p=0.0020) as well as virus-negative only (p=0.15). Furthermore, the proportion of mutated residues in the helicase domain was significantly greater than that of the DEAD domain (p=0.0051, Fisher's exact test) and the region of the protein preceding the DEAD domain from amino acids 1-203 (p=0.0002, Fisher's exact test). The *DDX3X* gene is located on the X chromosome and was previously reported to escape X inactivation in females [260]. Both truncating events and at least some missense mutations in *DDX3X* have been previously described as causing functional loss of protein activity [261]. Notably, while only 60% of patients were male (910/1529 with available data), *DDX3X* mutations that were truncating occurred almost exclusively in males (19/20, 95%) (**Figure 4.4C**), consistent with a previous study [262]. This was

similarly evident in virus-positive (10/10, 100%) and virus-negative (9/10, 90%) cases separately. As 50% of patients with nonsynonymous *DDX3X* mutations were from the BL cohort (61/123 cases), we focused on this disease to evaluate the relationship between mutation status and DDX3X expression, using previously published RNA-sequencing data [249]. We found that male patients lacking somatic *DDX3X* mutations had a significantly lower expression of DDX3X compared to *DDX3X* unmutated female patients (median 13.79 and 14.50, p=2.4e-6, MWU test), consistent with escape from X inactivation [260]. Cases with missense mutations in *DDX3X* had an elevated expression of DDX3X irrespective of sex compared to cases with no mutations in *DDX3X* (median 15.04 and 14.24, p=1.2e-9), potentially suggesting that overexpression of missense mutants may favor their ability to decrease DDX3X function, while cases with truncating mutations (9 EBV-positive; 4 EBV-) had a lower expression (median 12.23 and 14.24, p=1.79e-5) (**Figure 4.4D**), consistent with them being loss-of-function events. Similarly, in the TCGA study of HNSCC, expression of DDX3X was lower in unmutated male cases compared to unmutated female cases (median 12.58 and 12.99, p<2.2e-16), and was also lower in cases (3 HPV+; 2 HPV-) with truncating mutations compared to unmutated cases (median 10.37 and 12.71, p=1.5e-4) (**Figure 4.S.6**). Together, these results suggest that mutations in *DDX3X* and *EIF4A1* may play a role in virus-positive tumors in various types of cancer (**Figure 4.4E**).

***HLA-I* homozygosity is a germline risk factor for EBV-positive Hodgkin lymphoma**

In light of the essential role of the major histocompatibility complex class I (MHC-I) in the adaptive immune response to viral infection and oncogenesis, we investigated the role of germline allele type of the three *HLA-I* genes (*HLA-A*, *HLA-B*, and *HLA-C*) encoding the heavy chain subunit of the MHC-I. Germline *HLA-I* typing of 1,255 patients representing 8 types of cancer revealed EBV-positive cHL had the highest rate of germline homozygosity in at least one *HLA-I*

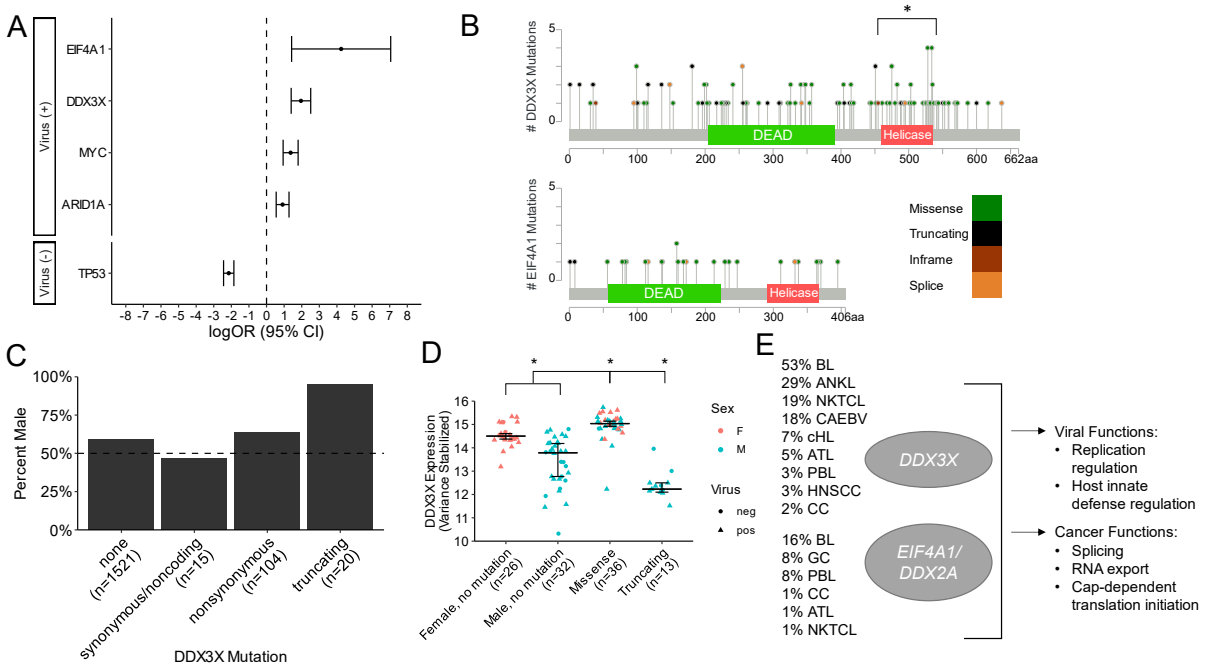


Figure 4.4: Somatic mutations in *EIF4A1* and *DDX3X* are recurrent genetic lesions associated with virus-positive status. A) Combined odds ratio of mutation in genes associated with virus-positive (top) and virus-negative (bottom) status (q<0.05) from pooled data of 1,658 tumors from 9 virus-associated cancers. *q < 0.05, Fisher's exact test, BH corrected. B) Mutations in *DDX3X* and *EIF4A1* in 1,974 tumors. * p < 0.05, binomial test. C) Fraction of patients that are male by *DDX3X* mutation status. D) DDX3X expression by *DDX3X* mutation status and sex in Burkitt lymphoma (n=117). * p < 0.05, MWU test. E) Frequencies of mutation of *DDX3X* and *EIF4A1* in virus-positive tumors overall and summary of key biological functions.

gene (8/15, 53%), which was significantly higher than expected based on the rate in the general population (21%) estimated from a subset of the GTeX database [263] (**Figure 4.5A**). Similar findings were seen in the UK BioBank [67], where 28% (156/562) of cHL cases (virus-positive and negative combined) were homozygous in at least one *HLA-I* locus, significantly higher than the normal U.K. population (23%) (**Figure 4.S.6**). These findings suggest germline *HLA-I* homozygosity may be an inherited risk factor for the development of EBV-positive cHL.

The cHL patients with germline homozygosity in *HLA-I* display unique phenotypic characteristics compared to those that are germline heterozygous. Germline homozygous patients in the DNA sequencing cohort were mostly male (9/16, 56%), and older than patients who were

heterozygous at *HLA-I* (median age 52.5 in homozygous, 28.5 in heterozygous; p=0.074, MWU test). Within the UK BioBank, cHL patients who were fully heterozygous exhibited a bimodal diagnosis age curve with a larger peak from age 20-30 and a smaller peak at age 60. In contrast, cHL patients who were male and fully homozygous in all three *HLA-I* loci exhibited a diagnosis age curve with the greatest peak at ages 50-60, following a different age distribution compared to heterozygotes (p=0.072, Kolmogorov-Smirnov [KS] test) (**Figure 4.5B**). These phenotypic characteristics are consistent with what has been observed clinically for EBV-associated cases and may reflect an enrichment of germline *HLA*-homozygous cases within the EBV-positive subtype.

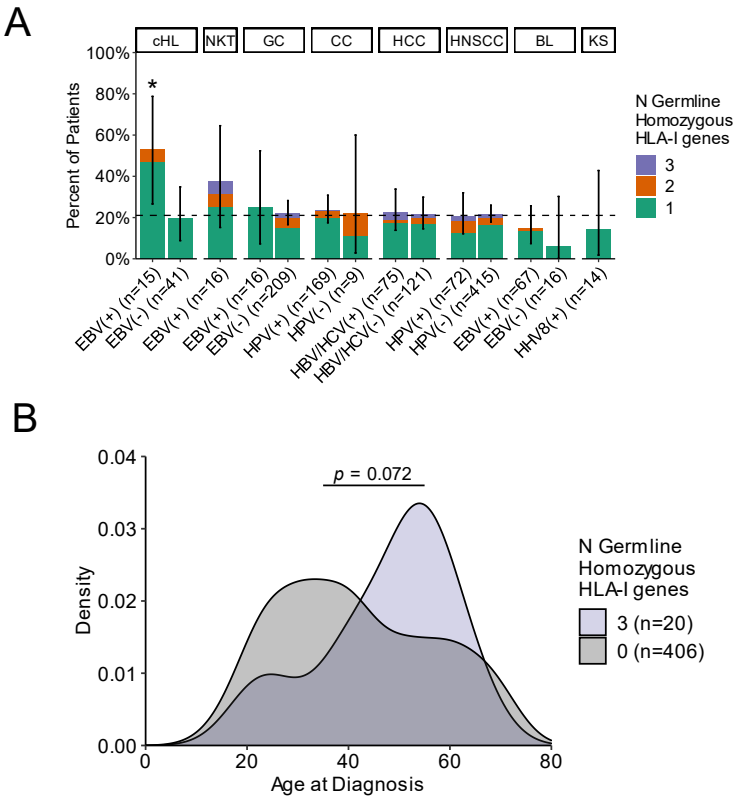


Figure 4.5: Germline *HLA-I* zygosity in virus-associated malignancies. A) Percent of germline homozygous individuals in *HLA-I* in eight virus-associated tumors by virus-infection status (positive versus negative). B) Distribution of age of cHL diagnosis in UK BioBank cHL patients that were germline heterozygous in *HLA-I* (n=406) versus germline homozygous in all three *HLA-I* genes (n=20).

Virus-associated cancers exhibit better responses to immunotherapy

PD-L1 overexpression has been linked to better overall survival in patients treated with immune checkpoint inhibitors (ICI) in several tumor types, including gastric cancer [264], head and neck cancers [265], and Merkel cell carcinomas [266]. PD-L1 expression has been associated with infection by oncoviruses including EBV [267], HPV [268], HBV [269], and MCPyV [266]. To determine whether virus positivity might be a useful marker for response to ICI therapy, we evaluated the correlation of viral status with response to ICI therapy with anti-PD(L)1 in 39 studies reported on ClinicalTrials.gov that had available therapy response and virus infection status data, representing four virus-linked cancers.

Virus positivity was significantly associated with ICI therapy response in GC (OR 2.27, 95% CI 1.17-4.29, p=0.011, Fisher's exact test) and HNSCC (OR 1.89, 95% CI 1.27-2.82, p=0.0012), but not MCC (OR 1.09, 95% CI 0.49-2.45, p=0.85) or HCC (OR 1.27, 95% CI 0.94-1.73, p=0.12) (**Figure 4.6A**). The same tumors displayed significant association between PD-L1 expression and ICI therapy response, including GC (OR 3.85, 95% CI 2.29-6.72, p<1.0e-5), HCC (OR 1.52, 95% CI 1.08-2.13, p=0.012), and HNSCC (OR 1.89, 95% CI 1.01-3.80, p=0.045), whereas a trend was observed in MCC (OR 2.25, 95% CI 0.76-7.63, p=0.15). As expected, higher tumor mutational burden (TMB) was associated with ICI therapy response in both GC (OR 3.55, 95% CI 2.09-6.08, p=7.74e-7) and HNSCC (OR 4.31, 95% CI 3.25-5.71, p<2.2e-16), the two cancer types with such data available.

In order to determine whether virus positivity is an independent marker of ICI therapy response, we compared the expression of PD-L1 [CD274] in TCGA's studies of GC [242], HCC [243], and HNSCC [245]. PD-L1 expression was higher in EBV-positive GC compared to EBV-negative GC (median 103.03 and 34.81, p=1.3e-7), but not HCC or HNSCC (**Figure 4.6B**). These

results suggest that EBV-positive status could be a positive prognostic marker for patients undergoing ICI therapy for gastric and head and neck cancers associated with EBV infection, which may be correlated with PD-L1 expression in GC but may represent an independent marker in HNSCC.

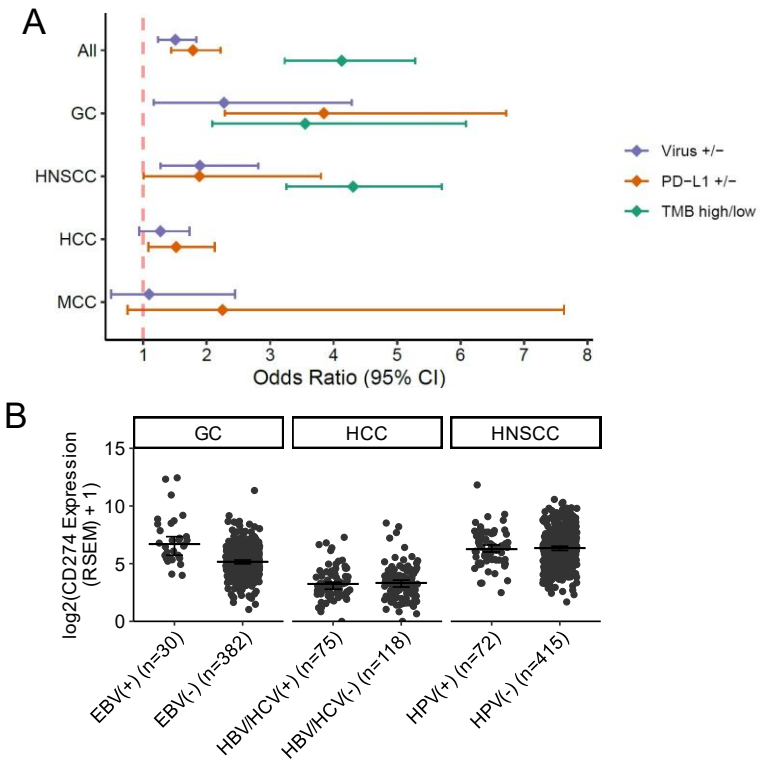


Figure 4.6: Meta-analysis of immunotherapy trials in virus-associated cancers. **A) Odds ratio of positive response to treatment with ICIs with virus-positive status, PD-L1 positive status, and/or high tumor mutation burden (TMB) in 37 studies representing four type of cancer. B) Expression of PD-L1 (CD274) versus viral status of tumors in TCGA's studies of GC (TCGA-STAD), HCC (TCGA-LIHC), and HNSCC (TCGA-HNSC).**

4.3 Discussion

This study provides insights into the epidemiological, inherited, somatic, and immune components commonly implicated in the pathogenesis of cancers associated with oncoviruses. Through analysis of cancer incidence rates reported in a selection of published studies, we noted virus-associated cancers display greater incidence in males compared to females relative to non-virus-associated cancers. The greater incidence of virus-associated cancers in males may be caused

in part by immunologic predisposition towards viral infection compared to females. For example, males infected with HBV are more likely to become viral carriers, whereas females infected with HBV are more likely to develop antibodies indicative of recovery and immunity to the virus [270]. In general, females have a more robust immune response to infection than males, which has been attributed to X-chromosome inactivation and regulation of the immune response by genetic, hormonal, and environmental mediators [271, 272].

Through a large-scale analysis of DNA sequencing data from 1,658 tumors collected from different studies, we found that virus-positive tumors generally display a lower mutation load compared to virus-negative tumors. It has been hypothesized that the oncogenic activity of virus-encoded proteins removes selective pressure for somatic mutations. For example, in cHL, the rarity of mutations in the PI3K-AKT signaling inhibitor *GNA13* in EBV-positive cases may be explained by the activity of LMP2a, which has been shown to activate the PI3K-AKT pathway [106]. However, unlike the other virus-associated cancers, virus-positive hepatocellular carcinomas and Burkitt lymphomas (i.e. those confirmed EBV-positive, comprised mostly but not exclusively of endemic BL) have a greater mutation load compared to virus-negative cases. In HCC, this may reflect increased genomic instability of virus-positive HCC tumors resulting from integration of HBV into the host cell genome [273, 274], and/or the activity of HCV oncoproteins that inhibit DNA repair and induce double-stranded breaks in the DNA [275]. The greater genome-wide mutation load in EBV-associated BL has been attributed, at least in part, to the presence of mismatch repair and *AICDA*-mediated ASHM signatures in these cases, though the mechanism of mutagenesis by EBV in BL is yet to be elucidated [249]. Yet, EBV-positive BL had a lower driver mutation load compared to EBV-negative BL, which may be attributed to the activity of viral proteins such as EBNA1 that reduce selective pressure for drivers genetic lesion seen commonly

in virus-negative Burkitt lymphomas [249]. Interestingly, we found evidence for decreased ASHM in EBV-positive cHL compared to EBV-negative cHL, showing the opposite trend from that seen in BL. This finding highlights how mutation processes may differ even in cancers associated with the same virus, potentially related to the translation of distinct viral oncoproteins from different viral latency programs.

We found that somatic mutation of the RNA helicase gene *DDX3X* was more frequent in virus-positive tumors compared to virus-negative tumors overall and for a variety of individual cancer types. *DDX3X* is a member of the DEAD (Asp-Glu-Ala-Asp) box protein family involved in multiple functions related to RNA metabolism, including transcription regulation, splicing, RNA export, and translation initiation [260]. *DDX3X* additionally functions as a component of the innate immune signaling pathway, and is known to inhibit replication of viruses such as HBV by activating production of IFN-beta [260, 276]. Some RNA viruses, including HCV and HIV, exploit functions of *DDX3X* to aid in viral replication [260, 276]. In cancer, *DDX3X* has been described as both a tumor suppressor and an oncogene in different cancer types and even among different tumors of the same cancer type [277]. *DDX3X* is expressed in many tissues of the body and escapes X-chromosome inactivation [260, 277]. The relatively high frequency of mutations in *DDX3X* in virus-positive tumors and the near-exclusive male bias for truncating mutations suggests loss of function of *DDX3X* may contribute to the pathogenesis of some virus-associated cancers, particularly Burkitt lymphoma, which had the highest frequency of *DDX3X* mutations in this study and for which similar findings were recently reported in another study [262].

The greater frequency of *TP53* mutations in virus-negative compared to virus-positive tumors found in this analysis may reflect how viral oncoproteins inhibit the activity of tumor suppressors. EBV-encoded oncoproteins have been shown to inhibit tumor suppressive functions

of p53 and other proteins in the p53 family [278]. Similar functions have been observed in other herpesviruses, including HHV8 (associated with KS) [278]. The E6 and E7 proteins encoded by HPV inhibit p53 and Rb, respectively [148]. In these and other cancers, the lower frequency of *TP53* mutations may reflect the lack of selective pressure for *TP53* mutations due to the disruption of p53 functions by viral oncoproteins.

The results of our study indicate the immune response plays a critical role in the risk, development, and/or response to therapy of virus-positive tumors. Alterations in MHC-I were found to follow a trend in cHL specifically, where somatic inactivating events were more frequent in virus-negative compared to virus-positive tumors, possibly to limit the presentation of neo-epitopes generated by the higher mutation burden in EBV-negative cHL, whereas germline homozygosity of *HLA-I* was more frequent in EBV-positive cHL, possibly to limit the presentation of viral antigens. Consistent with this finding, cHL had an elevated rate of germline *HLA-I* homozygosity compared to normal individuals in the UK BioBank, and cHL patients homozygous in all three *HLA-I* loci tended to be older patients, a group known to be enriched for EBV-positivity compared to young adult cHL. More frequent germline *HLA-I* homozygosity in EBV-positive cHL as a means to reduce viral antigen presentation, and more frequent somatic *HLA-I* inactivation in EBV-negative cHL as a means to reduce tumor antigen presentation, may contribute to explain the observations that normal *HLA-I* surface expression by cHL tumor cells is largely preserved in EBV-positive cases and largely lost in EBV-negative cases [88], respectively. It remains to be explained why germline *HLA-I* homozygosity was not increased in other EBV-associated cancers such as EBV-positive BL [33], and why germline homozygosity represents a common feature of DLBCL, a largely non-virus-associated B-cell lymphoma [64]. One explanation may be that BL has a more restricted viral latency antigen expression compared to cHL [150]. In DLBCL, the high

rate of germline *HLA-I* homozygosity (26%, compared to 28% in HL and 23% in normal) was thought to contribute to limit neoantigen repertoire presentation, given the substantial non-synonymous mutation burden typical of this disease, including ASHM [33]. Although the mutational burden of EBV-negative cHL may appear higher than DLBCL overall [249], the phenotypic and molecular heterogeneity of the latter warrants further analyses focused on individual subtypes in larger number of cases. Future work should verify how germline *HLA-I* zygosity can predispose to the development of EBV-positive cHL.

Analysis of ICI clinical trials reveals that virus-positive status could represent a positive biomarker for ICI therapy response in GC and HNSCC. The improved response to immunotherapy of EBV-positive GC patients compared to EBV-negative GC patients is hypothesized to be due to increased expression of PD-L1, potentially through activation of the NF-κB pathway by viral protein LMP2A [279]. This is consistent with the association between PD-L1 expression and EBV-positive status in TCGA's study of GC, as well as other studies that reported similar results in GC tumors [267]. The association between HPV infection and PD-L1 expression is less clear: some studies report a link between HPV status and PD-L1 expression [280, 281], while others find no association [282, 283], the latter of which is consistent with the results from the TCGA HNSCC dataset. It has been hypothesized that the improved response of HPV-positive HNSCC tumors to ICI therapies may be due to increased abundance of tumor infiltrating lymphocytes and CD8+ T cells in the microenvironment of virus-positive tumors [284].

Our integrative analysis of the epidemiological and genetic factors related to virus-associated cancers highlights two approaches to oncogenesis, in the presence and absence of viral infection**.** In the absence of viral infection, a normal cell acquires somatic drivers towards malignant transformation through random mutations with age, defective DNA repair, exogenous

carcinogens, and interactions with the microbiome, together with selection. Following the acquisition of an average of 5-7 driver mutations [285], the normal cell transforms into a cancer cell (**Figure 4.7A**). In virus-positive cancers, a normal cell is first infected with a virus, potentially as a result of inherited or environmental risk factors. The infected cell acquires somatic mutations over time, though fewer are generally required compared to the non-viral scenario due to the activities of viral oncoproteins. Finally, after acquiring the requisite number of drivers, the infected

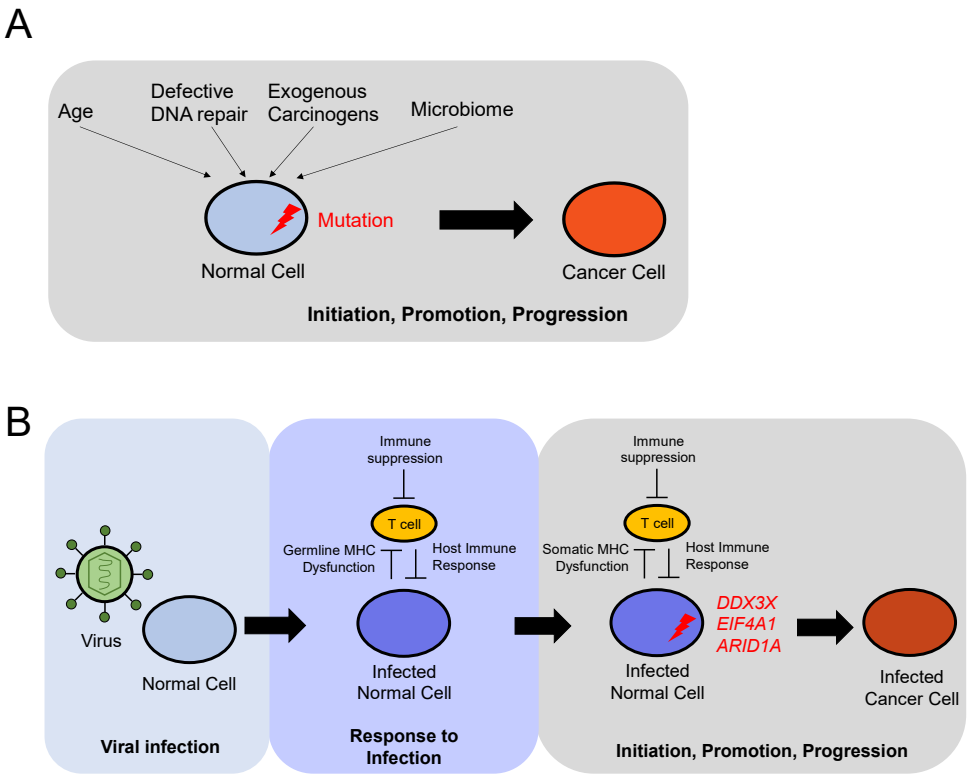


Figure 4.7. Example of oncogenesis in the presence and absence of viral infection. A) Schematic representation of oncogenesis in the absence of viral infection. A normal cell accumulates driver mutations as a result of age, defective DNA repair, exogenous carcinogens, or microbiome interactions, leading to initiation, promotion, and progression that ends in the malignant transformation of the cell. B) Schematic representation of oncogenesis in the presence of viral infection. Infection of normal cell is established as a result of inadequate host immune response, potentially associated with germline MHC dysfunction or other inherited risk factors. The infected normal cell acquires somatic mutations in specific genes, such as chromatin modifiers like *ARID1A* or RNA helicases *DDX3X* and *EIF4A1*, leading to initiation, promotion, and progression that ends in the malignant transformation of the infected cell.

cell transforms into a cancer cell, and continues to follow a trajectory of progression distinct from the non-viral scenario (**Figure 4.7B**). Further studies will be needed in order to understand how differences in the development and progression of tumors due to viral infection following this model may be incorporated into the development of targeted therapies.

4.4 Methods

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Overview

Clinical and genomic data of 1,974 cancer patients was obtained from 14 published studies [101, 107, 109, 242, 244-249, 256-259] (1,963 patients) and newly collected DNA sequencing data of Hodgkin lymphoma (32 patients) and Kaposi sarcoma (10 patients). The combined cohort consists of 612 females, 903 males, and 356 individuals of unknown or unreported sex. The ages range from 1 year to 90 years.

METHOD DETAILS

Epidemiological Analysis

Sex ratio in incidence rates of virus-associated and non-virus associated cancers, as well as virus-positive and virus-negative cases of virus-associated cancers, were obtained from 86 published studies [56, 156-236]. Global age-standardized incidence rates of cancers by country in 2020 were obtained from GLOBOCAN 2020 Cancer Today online portal (https://gco.iarc.fr/today/home). Attributable fraction of cancer cases for each region were obtained from de Martel et al [44].

Virus infection status calling

EBV infection status of cHL patients was determined by standard EBER in-situ hybridization on fixed tissue sections and confirmed by presence of reads aligned to the EBV

reference assessed by samtools idxstats. Virus-infection status of patients in other cohorts were reported in the original studies [93, 101, 107, 243, 246-249, 256-259] or obtained via cBioportal (https://www.cbioportal.org/) for TCGA cervical [244], gastric [242], and head and neck squamous cell carcinoma [245] data sets.

Single nucleotide and indel variant calling pipeline

Hodgkin lymphoma

Variant calling of cHL samples was performed as described in Chapter 2. Two cases (patient IDs c_cHL_4 and c_cHL_24) were excluded from mutation load comparison due to previously described microsatellite instability [107].

Kaposi sarcoma

Whole genome sequencing samples were aligned to GRCh37 using the Burrows-Wheeler aligner. Samples were pre-processed by indel realignment, duplicate removal, and base recalibration with GATK [120] following the GATK best practices workflow [121]. SAVI-v2 [122], an in-house variant caller, was used to call somatic variants. The variant list was filtered for variants with a minimum total depth 10 and maximum total depth 700 in both tumor and normal, strand bias p value > 0.001 in tumor and normal, and called as significant somatic variants by SAVI (p-value <0.05, and confidence interval for the significance of the tumor/normal comparison >0). Variants were excluded if they were found in an in-house supernormal created from 186 normal samples from the TCGA, if they were in the cohort supernormal constructed from variants in the ten normal samples, or if they were common SNPs found at a frequency >= 5% in the 1000 Genomes Project.

Mutation signature analysis

Mutation signatures were called from somatic variants separately for each cancer type

using Palimpsest [123], an NMF-based mutation signature caller. Signatures were obtained from somatic variants called from whole genome sequencing data when available (cHL) or whole exome sequencing data (other cancers). First, unsupervised mutation signature analysis was run for single base substitution (SBS), double base substitution (DBS), and insertion-deletion (ID) signature calling (when applicable) for all variants in all patients of a given tumor type. The similarity between *de novo* inferred signatures and published mutation signatures from COSMIC v3 was assessed by cosine similarity function in Palimpsest. Supervised mutation signature analysis was then run separately for each cancer type with mutation signatures that were the highest ranking match for the inferred signature from the unsupervised analysis. Mutation signature calling in cHL samples followed the methods described in Chapter 2.

Copy number segmentation and variant calling

Copy number segmentation of Kaposi sarcoma samples was conducted using Sequenza [131] for each pair of tumor of normal samples sequenced by whole exome sequencing for each case. Copy number segmentation of plasmablastic lymphoma samples was performed with Oncoscan, as previously described [101]. Copy number segmentation data of other cancers was obtained from the original published studies (Burkitt [249], NKTCL [258]) or cBioportal (TCGA samples [242-245]).

GISTIC Analysis

In order to define significant regions of recurrent CNAs across virus-associated cancers, GISTIC 2.0 [130] was applied to pooled copy number segmentation data of 1,557 tumors, using a significance threshold of q = 0.1, a maximum segmentation threshold of 10,000, and all other parameters default via the GenePattern server (https://www.genepattern.org/). GISTIC peaks of gain or amplification were counted as present in a patient if the maximum inferred tumor copy

number within the wide peak limit region was > 2.3 (gain) or > 3.6 (amplification). GISTIC peaks of deletion were counted as present in a patient if the minimum inferred tumor copy number within the wide peak limit region was < 1.7 (heterozygous loss) or < 0.8 (homozygous loss). Arm and whole-chromosome level CNAs were defined as lesions of the same type (i.e. gain or loss) that covered >75% of the chromosome arm or chromosome, respectively.

Analysis of Recurrently Mutated Genes in Virus-associated Cancers

Nonsynonymous mutations in protein-coding genes were counted in 1,658 patients with DNA sequencing data. For patients with both WES and WGS available, WES was used due to greater depth of sequencing. Hodgkin lymphoma cases were filtered to include only clonal mutations (Tfreq >=20 in both tumor samples) to reduce noise from the whole genome amplification procedure. Cases of MCC and PCNSL were only counted in the statistics for genes which were included in those targeted panels. To eliminate noise from hypermutated samples, for all cancers, only cases with <300 mutations were included. Significant genes were defined as those with an OR > 1 and BH corrected p value < 0.05 and recurrently mutated in at least two cases in at least two unique cancer types.

***HLA-I* Analysis**

Molecular Data Sets

Class I *HLA* allele typing of DNA sequencing samples (56/56 cHL [109], 10 newly sequenced Kaposi sarcoma, 93 BL [249]) was performed using PolySolver with default parameters. Class I *HLA* allele typing of 4 additional previously published Kaposi sarcoma samples [286] and all NKTCL samples [258] from RNA sequencing was performed using arcasHLA [287] with default parameters. Class I *HLA* typing of TCGA data sets from RNA-seq was obtained from a previous study [288]. We counted as “homozygous” those cases where both

inferred alleles of an *HLA-I* gene are the same to the two-field resolution (allele group and specific HLA protein).

UK BioBank

Imputed *HLA-I* genotypes of 488,265 patients from the UK BioBank reported by HLA*IMP:02 were obtained as described in ref. [33]. UK BioBank individuals were categorized into different cancer groups by organ location/pathology according to hospital and cancer registry records using ICD10 codes. The rate of homozygosity in the general population was estimated from RNA-seq of 95 samples representing 5 tissue types in the GTEx data base, as previously described [33].

Analysis of Immunotherapy Trials

The comparative analysis of response to immunotherapy was performed using data from ClinicalTrials.gov. Eleven checkpoint inhibitors targeting either PD-1, PD-L1, or CTLA-4 were included: Ipilimumab, Nivolumab, Pembrolizumab, Cemiplimab, Atezolizumab, Avelumab, Durvalumab, Camrelizumab, Sintilimab, Toripalimab, Tremelimumab. On September 22nd 2022, trials were collected using the following query: "((NOT NOTEXT) [CITATIONS]) AND (<ICI drug name 1> OR <ICI drug name 2> OR …)", where ICI drug names correspond to the ones listed above as well as any known synonyms (e.g. Ipilimumab: BMS-734016, MDX-010, MDX-101, Yervoy). Only single-arm, interventional studies where the objective response rate was available specifically for immune checkpoint therapy (with no other combination therapies) were included. Whenever possible, response stratified by virus status, PD-L1 expression and tumor mutational burden was collected.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analyses of significance of mutation counts and frequencies were performed using a

Mann-Whitney U test and two-sided Fisher's exact test, respectively. Significance of mutation signature activities were assessed using Student's t test. Odds ratios of mutation by virus status were computed with Haldane-Anscombe correction when applicable. The 95% confidence interval of odds ratios is reported as the normal approximation (Wald). The distribution of *HLA-I* germline homozygosity in HL patients in the UK BioBank was compared using a Kolmogorov-Smirnov test. Multiple hypothesis corrections were applied using the Benjamini-Hochberg Procedure and reported as q-values.

4.5 Supplementary Figures

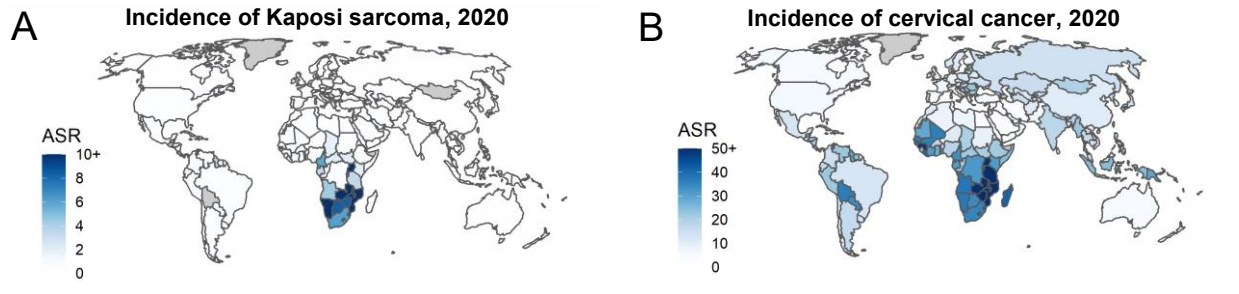


Figure 4.S.1. Geographic distributions of Kaposi sarcoma and cervical cancer by country reported by GLOBOCAN 2020. A) Estimated age-standardized incidence rate (ASR) of Kaposi sarcoma by country. B) Estimated ASR of cervical cancer by country.

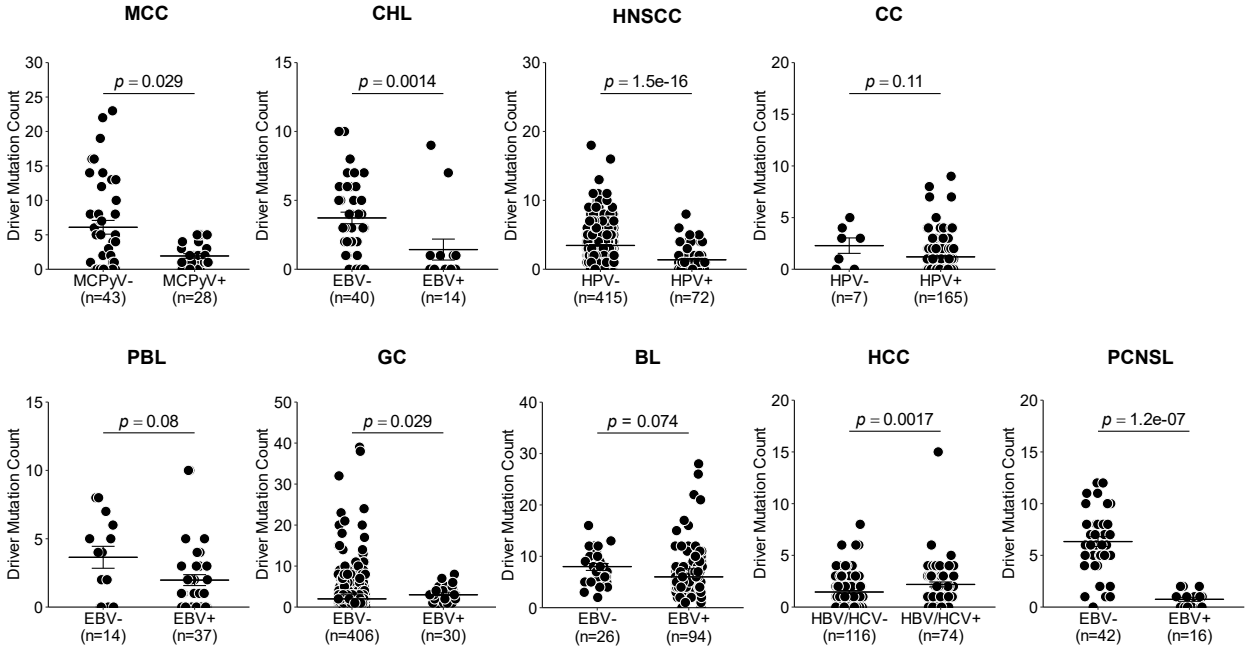


Figure 4.S.2. Count of mutations in driver genes in virus-positive versus virus-negative tumors in nine cancers. Driver genes are defined as those genes listed as recurrently mutated and/or known drivers associated with the specific cancer in the original study.

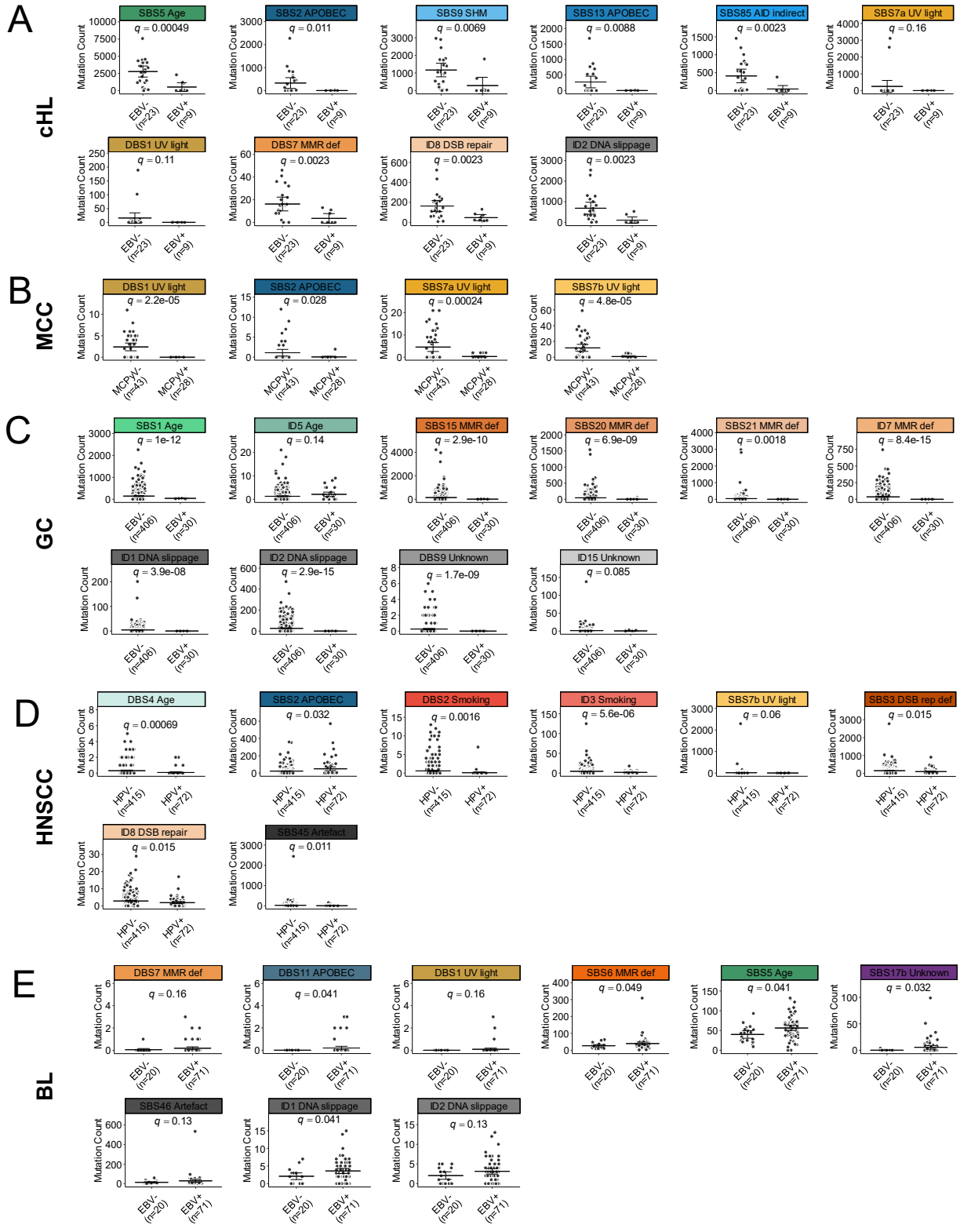


Figure 4.S.3. Counts of mutations attributed to each mutation signature in virus-positive and virus-negative cases of five cancers. A) cHL (n=32); B) MCC (n=71); C) GC (n=436); D)

HNSCC (n=487); E) BL (n=91). Comparisons with a q < 0.2 are shown. q values from Student's t test, BH corrected.

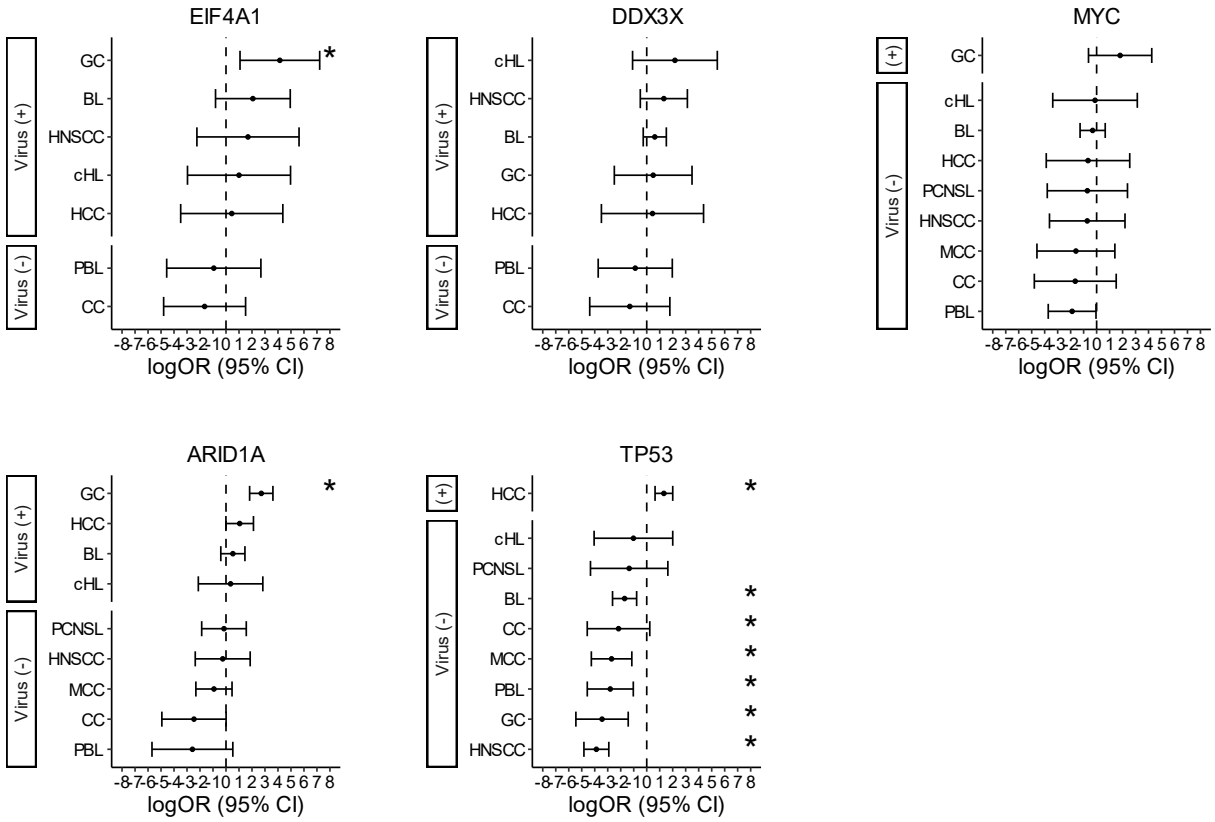


Figure 4.S.4. Odds ratio of mutation in virus-positive versus virus-negative tumors by cancer type. Genes with a significant OR of mutation in virus-positive tumors from the combined cohort are shown. * q < 0.05, chi-square test, BH corrected.

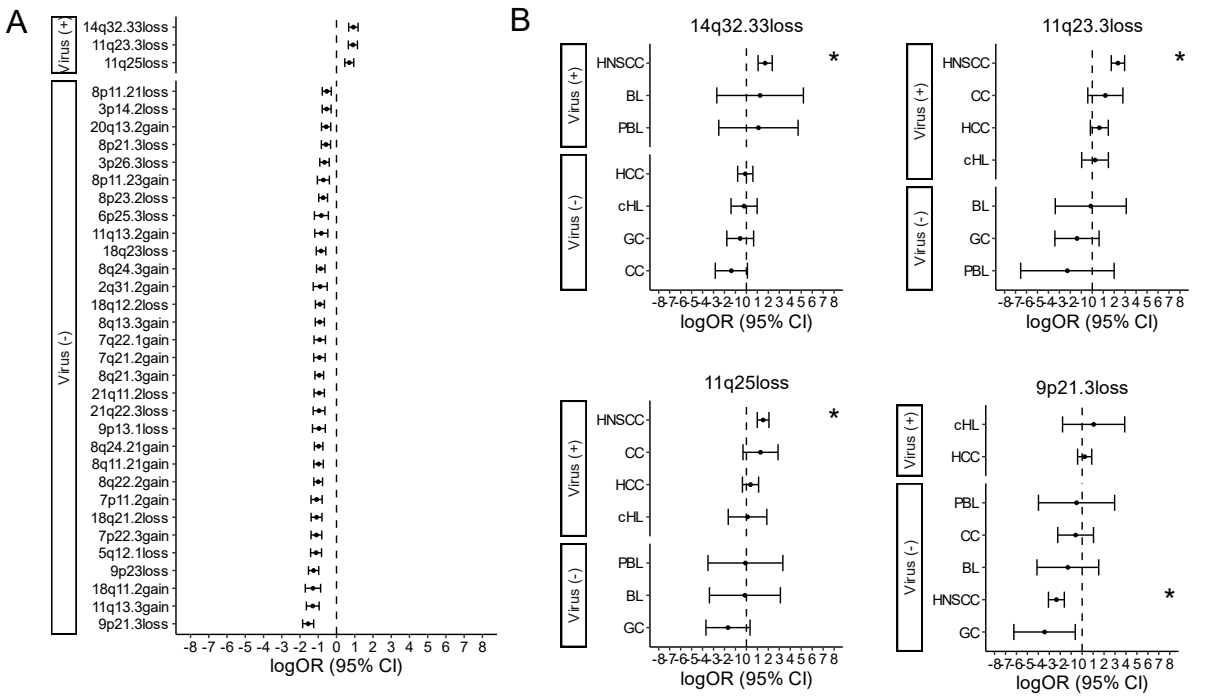


Figure 4.S.5. Recurrent copy number alterations (CNAs) in virus-associated cancers. A) Odds ratio of CNA in virus-positive versus virus-negative tumors in the combined cohort. B) Odds ratio of CNA in virus-positive versus virus-negative tumors by cancer type in regions with a significant OR of CNA in virus-positive tumors from the combined cohort and the top ranking region with a significant OR of CNA in virus-negative tumors from the combined cohort. * q < 0.0001, chi-square test, BH corrected.

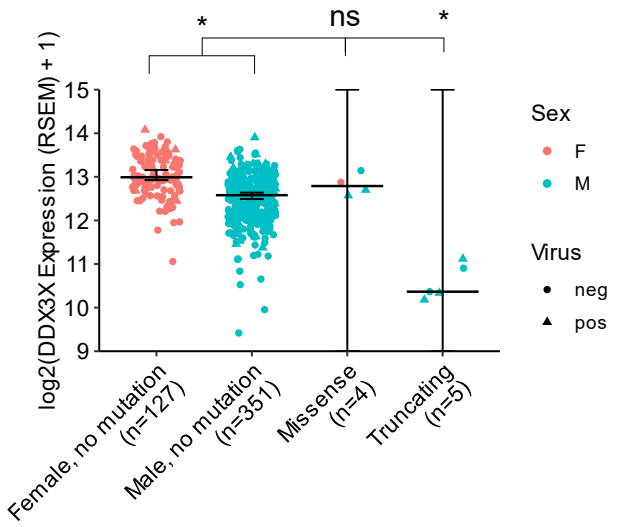


Figure 4.S.6. Expression of *DDX3X* and *DDX3X* mutation status in TCGA-HNSC (n=487). * p < 0.05, MWU test.

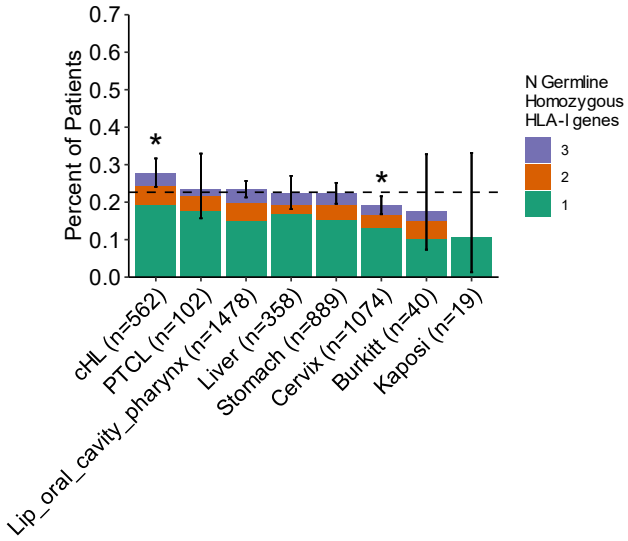


Figure 4.S.7. Frequency of homozygosity in *HLA-I* in virus-associated cancers in the UK BioBank. cHL: classical Hodgkin lymphoma; PTCL: peripheral T-cell lymphoma. * p < 0.05, binomial test.

Conclusions

In this thesis, we have analyzed the role of dysregulation of MHC-I in B-cell lymphomas, with a focus on DLBCL and cHL. The MHC-I complex is a key part of the adaptive immune response, and disruption of MHC-I functions through either inherited or somatic means is one way that tumor cells evade the immune system. We find that DLBCL and cHL are characterized by frequent somatic alterations of MHC-I as well as germline homozygosity of *HLA-I* genes.

In Chapter 2, we studied the relationship between *HLA-I* zygosity and overall survival in DLBCL within molecular subtypes. We performed *HLA-I* allele typing and survival analysis on 519 DLBCL patients previously classified by genetic subtype on the basis of their somatic lesions. Patients whose tumors belonged to the EZB subtype and who were homozygous in at least one *HLA-I* gene had a significantly worse overall survival compared to EZB patients that were fully heterozygous. These patients were characterized by frequent somatic mutation in the *EZH2* gene, which has been linked with downregulation of MHC-I expression through epigenetic silencing. This analysis indicates a potential role for *HLA-I* allele type and zygosity in overall survival trends in DLBCL patients that differs by molecular subtype.

In Chapter 3, we analyzed the genetics of cHL tumors that are linked to infection with EBV in comparison to EBV-negative tumors. We found that EBV-positive cHL is genetically distinct from EBV-negative cHL and is characterized by low somatic mutation burden and unique patterns of mutation. Through *HLA* allele typing from normal DNA of these patients, we found that the rate of germline *HLA-I* homozygosity was greater in EBV-associated cHL compared to non-EBV-associated cHL. Somatic lesions in MHC-I, including mutations in *B2M*, *HLA-I*, and loss of heterozygosity of *HLA-I* were more frequent in EBV-negative cHL compared to EBV-positive cHL. These findings suggest germline *HLA-I* zygosity may contribute to the risk of development

of EBV-positive cHL and that somatic alterations of MHC-I in cHL differ based on viral infection status.

In Chapter 4, we characterized the genomic landscape of nine cancers associated with five different viruses. We noted that virus-associated cancers are characterized by distinct epidemiological trends, including a greater incidence in males and geographical disparities in incidence. An analysis of sequencing data from 1,658 tumors in nine types of cancer revealed virus-positive tumors generally have a lower somatic mutation load, different mutation signature activities, and recurrent mutations in RNA helicases *DDX3X* and *EIF4A1* compared to virus-negative tumors of the same tissue type. Through *HLA-I* allele typing of 1,255 patients, we found that EBV-positive cHL had an elevated rate of germline homozygosity in *HLA-I* alleles compared to not only EBV-negative cHL but also the other common virus-associated cancers. These results indicate that *HLA-I* homozygosity is a risk factor for EBV-positive cHL specifically and is unique among other virus-associated malignancies.

Further work will be needed to describe the mechanisms by which *HLA-I* germline status contributes to the risk of development of DLBCL and cHL. The effect of germline *HLA-I* status on response to treatment with immune checkpoint inhibitors in DLBCL and cHL also requires further study. Finally, more investigations will be needed to better understand the role of MHC-I germline status and/or somatic dysregulation in rare lymphomas where there is emerging evidence of MHC-I involvement, such as plasmablastic lymphoma (see Appendix A).

References

1. Abbas, A.K., A.H. Lichtman, and S. Pillai, *Antigen Presentation to T Lymphocytes and the Function of Major Histocompatibility Complex Molecules*, in *Cellular and Molecular Immunology*. 2022, Elsevier: Philadelphia, Pennsylvania. p. 123-150.

2. Cruz-Tapias, P., J. Castiblanco, and J.-M. Anaya, *Major histocompatibility complex: antigen processing and presentation*, in *Autoimmunity: From Bench to Bedside [Internet]*, J.-M. Anaya, et al., Editors. 2013, El Rosario University Press: Bogota, Colombia.

3. Lev, A., et al., *Compartmentalized MHC class I antigen processing enhances immunosurveillance by circumventing the law of mass action.* Proc Natl Acad Sci U S A, 2010. 107(15): p. 6964-9.

4. Yewdell, J.W., E. Reits, and J. Neefjes, *Making sense of mass destruction: quantitating MHC class I antigen presentation.* Nat Rev Immunol, 2003. 3(12): p. 952-61.

5. Montealegre, S. and P.M. van Endert, *Endocytic Recycling of MHC Class I Molecules in Non-professional Antigen Presenting and Dendritic Cells.* Front Immunol, 2018. 9: p. 3098.

6. Hewitt, E.W., *The MHC class I antigen presentation pathway: strategies for viral immune evasion.* Immunology, 2003. 110(2): p. 163-9.

7. Leone, P., et al., *MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells.* J Natl Cancer Inst, 2013. 105(16): p. 1172-87.

8. Bell, M.J., et al., *The peptide length specificity of some HLA class I alleles is very broad and includes peptides of up to 25 amino acids in length.* Mol Immunol, 2009. 46(8-9): p. 1911-7.

9. Wooldridge, L., et al., *A single autoimmune T cell receptor recognizes more than a million different peptides.* J Biol Chem, 2012. 287(2): p. 1168-77.

10. Alberts, B., et al., *T cells and MHC proteins*, in *Molecular Biology of the Cell. 4th edition*. 2002, Garland Science: New York.

11. Greenfield, E.A., K.A. Nguyen, and V.K. Kuchroo, *CD28/B7 costimulation: a review.* Crit Rev Immunol, 1998. 18(5): p. 389-418.

12. Swann, J.B. and M.J. Smyth, *Immune surveillance of tumors.* J Clin Invest, 2007. 117(5): p. 1137-46.

13. Male, D., R.S. Peebles, and V. Male, *Immunity to Cancers*, in *Immunology*. 2021, Elsevier. p. 319-323.

14. Morgan, D.J., et al., *Activation of low avidity CTL specific for a self epitope results in tumor rejection but not autoimmunity.* The Journal of Immunology, 1998. 160(2): p. 643-651.

15. Lee, C.-H., et al., *Update on tumor neoantigens and their utility: why it is good to be different.* Trends in immunology, 2018. 39(7): p. 536-548.

16. Gross, L., *Intradermal immunization of C3H mice against a sarcoma that originated in an animal of the same line.* Cancer research, 1943. 3(5): p. 326-333.

17. Matsushita, H., et al., *Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting.* Nature, 2012. 482(7385): p. 400-4.

18. Cornel, A.M., I.L. Mimpen, and S. Nierkens, *MHC Class I Downregulation in Cancer: Underlying Mechanisms and Potential Targets for Cancer Immunotherapy.* Cancers (Basel), 2020. 12(7).

19. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. 144(5): p. 646-74.

20. Genome Reference Consortium. *Genome reference consortium human build 38 patch release 14 (GRCh38. p14).* NCBI https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40/ 2019.

21. Allen, R.L. and L. Hogan, *Non-classical MHC class I molecules (MHC-Ib)*, in *eLS*. 2013, John Wiley & Sons, Ltd.: Chichester.

22. Goodfellow, P., et al., *The β2-microglobulin gene is on chromosome 15 and not in the HL-A region.* Nature, 1975. 254(5497): p. 267-269.

23. Trowsdale, J. and J.C. Knight, *Major histocompatibility complex genomics and human disease.* Annu Rev Genomics Hum Genet, 2013. 14: p. 301-23.

24. Jin, Y., et al., *Architecture of polymorphisms in the human genome reveals functionally important and positively selected variants in immune response and drug transporter genes.* Hum Genomics, 2018. 12(1): p. 43.

25. Robinson, J., et al., *IPD-IMGT/HLA Database.* Nucleic Acids Res, 2020. 48(D1): p. D948-D955.

26. Li, L., M. Dong, and X.G. Wang, *The Implication and Significance of Beta 2 Microglobulin: A Conservative Multifunctional Regulator.* Chin Med J (Engl), 2016. 129(4): p. 448-55.

27. Doherty, P.C. and R.M. Zinkernagel, *Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex.* Nature, 1975. 256(5512): p. 50-2.

28. Parham, P. and T. Ohta, *Population biology of antigen presentation by MHC class I molecules.* Science, 1996. 272(5258): p. 67-74.

29. Chowell, D., et al., *Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy.* Nat Med, 2019. 25(11): p. 1715-1720.

30. Grantham, R., *Amino acid difference formula to help explain protein evolution.* Science, 1974. 185(4154): p. 862-4.

31. Prugnolle, F., et al., *Pathogen-driven selection and worldwide HLA class I diversity.* Curr Biol, 2005. 15(11): p. 1022-7.

32. Abed, A., et al., *Prognostic value of HLA-I homozygosity in patients with non-small cell lung cancer treated with single agent immunotherapy.* J Immunother Cancer, 2020. 8(2): p. e001620.

33. Fangazio, M., et al., *Genetic mechanisms of HLA-I loss and immune escape in diffuse large B cell lymphoma.* Proc Natl Acad Sci U S A, 2021. 118(22): p. e2104504118.

34. Zawadzka-Starczewska, K., et al., *Actual Associations between HLA Haplotype and Graves' Disease Development.* J Clin Med, 2022. 11(9): p. 2492.

35. Gregersen, P.K., et al., *Risk for myasthenia gravis maps to a 151Pro→ Ala change in TNIP1 and to human leukocyte antigen-B* 08.* Annals of neurology, 2012. 72(6): p. 927-935.

36. Prinz, J.C., *Human Leukocyte Antigen-Class I Alleles and the Autoreactive T Cell Response in Psoriasis Pathogenesis.* Front Immunol, 2018. 9: p. 954.

37. Chen, B., et al., *Role of HLA-B27 in the pathogenesis of ankylosing spondylitis (Review).* Mol Med Rep, 2017. 15(4): p. 1943-1951.

38. Blackwell, J.M., S.E. Jamieson, and D. Burgner, *HLA and infectious diseases.* Clin Microbiol Rev, 2009. 22(2): p. 370-85.

39. Matzaraki, V., et al., *The MHC locus and genetic susceptibility to autoimmune and infectious diseases.* Genome Biol, 2017. 18(1): p. 76.

40. Miao, F., et al., *Association of human leukocyte antigen class I polymorphism with spontaneous clearance of hepatitis B surface antigen in Qidong Han population.* Clin Dev Immunol, 2013. 2013: p. 145725.

41. Carrington, M., et al., *HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage.* Science, 1999. 283(5408): p. 1748-52.

42. van Elsland, D. and J. Neefjes, *Bacterial infections and cancer.* EMBO Rep, 2018. 19(11): p. e46632.

43. Zapatka, M., et al., *The landscape of viral associations in human cancers.* Nat Genet, 2020. 52(3): p. 320-330.

44. de Martel, C., et al., *Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis.* Lancet Glob Health, 2020. 8(2): p. e180-e190.

45. Kapatai, G. and P. Murray, *Contribution of the Epstein Barr virus to the molecular pathogenesis of Hodgkin lymphoma.* J Clin Pathol, 2007. 60(12): p. 1342-9.

46. 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. 3310-5.

47. Jones, K., et al., *HLA class I associations with EBV+ post-transplant lymphoproliferative disorder.* Transplant Immunology, 2015. 32(2): p. 126-130.

48. Cerhan, J.R., et al., *Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma.* Nature genetics, 2014. 46(11): p. 1233-1238.

49. Falk, J.A. and D. Osoba, *The association of the human histocompatibility system with Hodgkin's disease.* International Journal of Immunogenetics, 1974. 1(1): p. 53-61.

50. Wang, S.S., et al., *Human leukocyte antigen class I and II alleles in non-Hodgkin lymphoma etiology.* Blood, 2010. 115(23): p. 4820-3.

51. Bosch, F.X., et al., *The causal relation between human papillomavirus and cervical cancer.* Journal of clinical pathology, 2002. 55(4): p. 244-265.

52. Dong, H., et al., *Current status of human papillomavirus-related head and neck cancer: from viral genome to patient care.* Virologica Sinica, 2021: p. 1-19.

53. Maucort-Boulch, D., et al., *Fraction and incidence of liver cancer attributable to hepatitis B and C viruses worldwide.* Int J Cancer, 2018. 142(12): p. 2471-2477.

54. Magrath, I.T., *African Burkitt's lymphoma. History, biology, clinical features, and treatment.* The American journal of pediatric hematology/oncology, 1991. 13(2): p. 222-246.

55. Hummel, M., et al., *Epstein-Barr virus in B-cell non-Hodgkin's lymphomas: unexpected infection patterns and different infection incidence in low-and high-grade types.* The Journal of pathology, 1995. 175(3): p. 263-271.

56. Castillo, J.J., M. Bibas, and R.N. Miranda, *The biology and treatment of plasmablastic lymphoma.* Blood, 2015. 125(15): p. 2323-30.

57. Brandsma, D. and J.E.C. Bromberg, *Primary CNS lymphoma in HIV infection.* Handb Clin Neurol, 2018. 152: p. 177-186.

58. Liu, Z., et al., *Characterization of the humoral immune response to the EBV proteome in extranodal NK/T-cell lymphoma.* Sci Rep, 2021. 11(1): p. 23664.

59. Murphy, G., et al., *Meta-analysis shows that prevalence of Epstein-Barr virus-positive gastric cancer differs based on sex and anatomic location.* Gastroenterology, 2009. 137(3): p. 824-33.

60. Tsao, S.W., C.M. Tsang, and K.W. Lo, *Epstein-Barr virus infection and nasopharyngeal carcinoma.* Philos Trans R Soc Lond B Biol Sci, 2017. 372(1732): p. 20160270.

61. Chang, Y., et al., *Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma.* Science, 1994. 266(5192): p. 1865-9.

62. Iwanaga, M., T. Watanabe, and K. Yamaguchi, *Adult T-cell leukemia: a review of epidemiological evidence.* Front Microbiol, 2012. 3: p. 322.

63. Yang, J.F. and J. You, *Merkel cell polyomavirus and associated Merkel cell carcinoma.* Tumour Virus Res, 2022. 13: p. 200232.

64. Hwang, J., et al., *The incidence of Epstein-Barr virus-positive diffuse large B-cell lymphoma: a systematic review and meta-analysis.* Cancers, 2021. 13(8): p. 1785.

65. Wang, S.S., et al., *HLA Class I and II Diversity Contributes to the Etiologic Heterogeneity of Non-Hodgkin Lymphoma Subtypes.* Cancer Res, 2018. 78(14): p. 4086-4096.

66. Goldin, L.R., et al., *Highly increased familial risks for specific lymphoma subtypes.* Br J Haematol, 2009. 146(1): p. 91-4.

67. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS Med, 2015. 12(3): p. e1001779.

68. Shankland, K.R., J.O. Armitage, and B.W. Hancock, *Non-hodgkin lymphoma.* The Lancet, 2012. 380(9844): p. 848-857.

69. Basso, K. and R. Dalla-Favera, *Germinal centres and B cell lymphomagenesis.* Nat Rev Immunol, 2015. 15(3): p. 172-84.

70. Moller, P., et al., *The primary mediastinal clear cell lymphoma of B-cell type has variable defects in MHC antigen expression.* Immunology, 1986. 59(3): p. 411-417.

71. Amiot, L., et al., *Loss of HLA molecules in B lymphomas is associated with an aggressive clinical course.* British journal of haematology, 1998. 100(4): p. 655-663.

72. Jordanova, E.S., et al., *β2-microglobulin aberrations in diffuse large B-cell lymphoma of the testis and the central nervous system.* International journal of cancer, 2003. 103(3): p. 393-398.

73. Pasqualucci, L., et al., *Analysis of the coding genome of diffuse large B-cell lymphoma.* Nat Genet, 2011. 43(9): p. 830-7.

74. Challa-Malladi, M., et al., *Combined genetic inactivation of beta2-Microglobulin and CD58 reveals frequent escape from immune recognition in diffuse large B cell lymphoma.* Cancer Cell, 2011. 20(6): p. 728-40.

75. Ennishi, D., et al., *Molecular and Genetic Characterization of MHC Deficiency Identifies EZH2 as Therapeutic Target for Enhancing Immune Recognition.* Cancer Discov, 2019. 9(4): p. 546-563.

76. Geiser, A.G., et al., *Transforming growth factor beta 1 (TGF-beta 1) controls expression of major histocompatibility genes in the postnatal mouse: aberrant histocompatibility antigen expression in the pathogenesis of the TGF-beta 1 null mouse phenotype.* Proceedings of the National Academy of Sciences, 1993. 90(21): p. 9944-9948.

77. Ma, D. and J. Niederkorn, *Transforming growth factor-beta down-regulates major histocompatibility complex class I antigen expression and increases the susceptibility of uveal melanoma cells to natural killer cell-mediated cytolysis.* Immunology, 1995. 86(2): p. 263.

78. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. 403(6769): p. 503-11.

79. Schmitz, R., et al., *Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma.* N Engl J Med, 2018. 378(15): p. 1396-1407.

80. Wright, G.W., et al., *A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications.* Cancer Cell, 2020. 37(4): p. 551-568 e14.

81. Chapuy, B., et al., *Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes.* Nat Med, 2018. 24(5): p. 679-690.

82. Lacy, S.E., et al., *Targeted sequencing in DLBCL, molecular subtypes, and outcomes: a Haematological Malignancy Research Network report.* Blood, 2020. 135(20): p. 1759-1771.

83. Morin, R.D., S.E. Arthur, and D.J. Hodson, *Molecular profiling in diffuse large B-cell lymphoma: why so many types of subtypes?* Br J Haematol, 2022. 196(4): p. 814-829.

84. Kaseb, H. and H.M. Babiker, *Hodgkin Lymphoma*, in *StatPearls [Internet]*. 2022, StatPearls Publishing.

85. Piris, M.A., L.J. Medeiros, and K.C. Chang, *Hodgkin lymphoma: a review of pathological features and recent advances in pathogenesis.* Pathology, 2020. 52(1): p. 154-165.

86. Vinnicombe, S.J., R.H. Reznek, and A. Rohatiner, *Chapter 29 - Hematologic Malignancy: The Lymphomas*, in *Oncologic Imaging: A Multidisciplinary Approach*, P.M. Silverman, Editor. 2012, W.B. Saunders: Philadelphia. p. 531-553.

87. Wong, Y., et al., *Estimating the global burden of Epstein-Barr virus-related cancers.* J Cancer Res Clin Oncol, 2022. 148(1): p. 31-46.

88. Nijland, M., et al., *HLA dependent immune escape mechanisms in B-cell lymphomas: Implications for immune checkpoint inhibitor therapy?* Oncoimmunology, 2017. 6(4): p. e1295202.

89. Oudejans, J., et al., *Analysis of major histocompatibility complex class I expression on Reed-Sternberg cells in relation to the cytotoxic T-cell response in Epstein-Barr virus-positive and-negative Hodgkin's disease.* Blood, 1996. 87(9): p. 3844-3851.

90. Murray, P.G., et al., *Analysis of major histocompatibility complex class I, TAP expression, and LMP2 epitope sequence in Epstein-Barr virus–positive Hodgkin's disease.* Blood, The Journal of the American Society of Hematology, 1998. 92(7): p. 2477-2483.

91. Lee, S.P., et al., *Antigen presenting phenotype of Hodgkin Reed-Sternberg cells: analysis of the HLA class I processing pathway and the effects of interleukin-10 on Epstein-Barr virus-specific cytotoxic T-cell recognition.* Blood, 1998. 92(3): p. 1020-1030.

92. Reichel, J., et al., *Flow sorting and exome sequencing reveal the oncogenome of primary Hodgkin and Reed-Sternberg cells.* Blood, 2015. 125(7): p. 1061-72.

93. Tiacci, E., et al., *Pervasive mutations of JAK-STAT pathway genes in classical Hodgkin lymphoma.* Blood, 2018. 131(22): p. 2454-2465.

94. Chowell, D., et al., *Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy.* Science, 2018. 359(6375): p. 582-587.

95. Chan, D.C., et al., *Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy.* Nature Medicine, 2019. 25(11): p. 1715-1720.

96. Lu, Z., et al., *Germline HLA-B evolutionary divergence influences the efficacy of immune checkpoint blockade therapy in gastrointestinal cancer.* Genome Med, 2021. 13(1): p. 175.

97. Lee, C.-H., et al., *High response rate and durability driven by HLA genetic diversity in kidney cancer patients treated with the immunotherapy combination lenvatinib and pembrolizumab.* Molecular cancer research, 2021. 19(9): p. 1510.

98. Ansell, S.M., et al., *Nivolumab for relapsed/refractory diffuse large B-cell lymphoma in patients ineligible for or having failed autologous transplantation: a single-arm, phase II study.* Journal of Clinical Oncology, 2019. 37(6): p. 481.

99. Mounier, N., et al., *Rituximab plus CHOP (R-CHOP) overcomes bcl-2—associated resistance to chemotherapy in elderly patients with diffuse large B-cell lymphoma (DLBCL).* Blood, 2003. 101(11): p. 4279-4284.

100. Coiffier, B. and C. Sarkozy, *Diffuse large B-cell lymphoma: R-CHOP failure-what to do?* Hematology Am Soc Hematol Educ Program, 2016. 2016(1): p. 366-378.

101. Liu, Z., et al., *Genomic characterization of HIV-associated plasmablastic lymphoma identifies pervasive mutations in the JAK-STAT pathway.* Blood Cancer Discov, 2020. 1(1): p. 112-125.

102. Alcoceba, M., et al., *HLA specificities are related to development and prognosis of diffuse large B-cell lymphoma.* Blood, 2013. 122(8): p. 1448-54.

103. Tada, K., et al., *Prognostic significance of HLA class I and II expression in patients with diffuse large B cell lymphoma treated with standard chemoimmunotherapy.* Cancer Immunol Immunother, 2016. 65(10): p. 1213-22.

104. Shukla, S.A., et al., *Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes.* Nat Biotechnol, 2015. 33(11): p. 1152-8.

105. Szolek, A., et al., *OptiType: precision HLA typing from next-generation sequencing data.* Bioinformatics, 2014. 30(23): p. 3310-6.

106. Mathas, S., S. Hartmann, and R. Kuppers, *Hodgkin lymphoma: Pathology and biology.* Semin Hematol, 2016. 53(3): p. 139-47.

107. Wienand, K., et al., *Genomic analyses of flow-sorted Hodgkin Reed-Sternberg cells reveal complementary mechanisms of immune evasion.* Blood Adv, 2019. 3(23): p. 4065-4080.

108. Spina, V., et al., *Circulating tumor DNA reveals genetics, clonal evolution, and residual disease in classical Hodgkin lymphoma.* Blood, 2018. 131(22): p. 2413-2425.

109. Tiacci, E., et al., *Pervasive mutations of JAK-STAT pathway genes in classical Hodgkin lymphoma.* Blood, 2018. 131(22): p. 2454-2465.

110. Alexandrov, L.B., et al., *Clock-like mutational processes in human somatic cells.* Nat Genet, 2015. 47(12): p. 1402-7.

111. Liso, A., et al., *Aberrant somatic hypermutation in tumor cells of nodular-lymphocyte-predominant and classic Hodgkin lymphoma.* Blood, 2006. 108(3): p. 1013-20.

112. Imielinski, M., G. Guo, and M. Meyerson, *Insertions and Deletions Target Lineage-Defining Genes in Human Cancers.* Cell, 2017. 168(3): p. 460-472 e14.

113. Pasqualucci, L., et al., *BCL-6 mutations in normal germinal center B cells: evidence of somatic hypermutation acting outside Ig loci.* Proc Natl Acad Sci U S A, 1998. 95(20): p. 11816-21.

114. Migliazza, A., et al., *Frequent somatic hypermutation of the 5' noncoding region of the BCL6 gene in B-cell lymphoma.* Proc Natl Acad Sci U S A, 1995. 92(26): p. 12520-4.

115. Seitz, V., et al., *Analysis of BCL-6 mutations in classic Hodgkin disease of the B- and T-cell type.* Blood, 2001. 97(8): p. 2401-5.

116. Liu, Z., et al., *Patterns of Human Leukocyte Antigen Class I and Class II Associations and Cancer.* Cancer Res, 2021. 81(4): p. 1148-1152.

117. Weniger, M.A. and R. Kuppers, *Molecular biology of Hodgkin lymphoma.* Leukemia, 2021. 35(4): p. 968-981.

118. Xia, Z., et al., *GNA13 regulates BCL2 expression and the sensitivity of GCB-DLBCL cells to BCL2 inhibitors in a palmitoylation-dependent manner.* Cell Death Dis, 2021. 12(1): p. 54.

119. Maura, F., et al., *Molecular Evolution of Classical Hodgkin Lymphoma Revealed Though Whole Genome Sequencing of Hodgkin and Reed-Sternberg Cells.* Blood, 2021. 138: p. 805.

120. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature genetics, 2011. 43(5): p. 491-498.

121. Van der Auwera, G.A., et al., *From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline.* Current protocols in bioinformatics, 2013. 43(1): p. 11.10. 1-11.10. 33.

122. Trifonov, V., et al., *SAVI: a statistical algorithm for variant frequency identification.* BMC Syst Biol, 2013. 7 Suppl 2: p. S2.

123. Shinde, J., et al., *Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer.* Bioinformatics, 2018. 34(19): p. 3380-3381.

124. Petljak, M., et al., *Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis.* Cell, 2019. 176(6): p. 1282-1294. e20.

125. Li, H., *Toward better understanding of artifacts in variant calling from high-coverage samples.* Bioinformatics, 2014. 30(20): p. 2843-51.

126. Pasqualucci, L., et al., *Genetics of follicular lymphoma transformation.* Cell Rep, 2014. 6(1): p. 130-40.

127. Okosun, J., et al., *Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma.* Nat Genet, 2014. 46(2): p. 176-181.

128. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.* Bioinformatics, 2012. 28(3): p. 423-5.

129. Chen, X., et al., *Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications.* Bioinformatics, 2016. 32(8): p. 1220-2.

130. Mermel, C.H., et al., *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol, 2011. 12(4): p. R41.

131. Favero, F., et al., *Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data.* Ann Oncol, 2015. 26(1): p. 64-70.

132. McGranahan, N., et al., *Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution.* Cell, 2017. 171(6): p. 1259-1271 e11.

133. Schrama, D., et al., *Merkel cell polyomavirus status is not associated with clinical course of Merkel cell carcinoma.* J Invest Dermatol, 2011. 131(8): p. 1631-8.

134. White, M.K., J.S. Pagano, and K. Khalili, *Viruses and human cancers: a long road of discovery of molecular paradigms.* Clin Microbiol Rev, 2014. 27(3): p. 463-81.

135. Epstein, M.A., B.G. Achong, and Y.M. Barr, *Virus Particles in Cultured Lymphoblasts from Burkitt's Lymphoma.* Lancet, 1964. 1(7335): p. 702-3.

136. Mundo, L., et al., *Frequent traces of EBV infection in Hodgkin and non-Hodgkin lymphomas classified as EBV-negative by routine methods: expanding the landscape of EBV-related lymphomas.* Mod Pathol, 2020. 33(12): p. 2407-2421.

137. de Martel, C., et al., *Worldwide burden of cancer attributable to HPV by site, country and HPV type.* Int J Cancer, 2017. 141(4): p. 664-670.

138. Narkhede, M., S. Arora, and C. Ujjani, *Primary effusion lymphoma: current perspectives.* Onco Targets Ther, 2018. 11: p. 3747-3754.

139. Calabro, M.L. and R. Sarid, *Human Herpesvirus 8 and Lymphoproliferative Disorders.* Mediterr J Hematol Infect Dis, 2018. 10(1): p. e2018061.

140. Feng, H., et al., *Clonal integration of a polyomavirus in human Merkel cell carcinoma.* Science, 2008. 319(5866): p. 1096-100.

141. Leiendecker, L., et al., *Human papillomavirus 42 drives digital papillary adenocarcinoma and elicits a germ-cell like program conserved in HPV-positive cancers.* Cancer Discov, 2022. CD-22-0489.

142. Poulin, D.L. and J.A. DeCaprio, *Is there a role for SV40 in human cancer?* J Clin Oncol, 2006. 24(26): p. 4356-65.

143. Siguier, M., P. Sellier, and J.F. Bergmann, *BK-virus infections: a literature review.* Med Mal Infect, 2012. 42(5): p. 181-7.

144. McLaughlin-Drubin, M.E. and K. Munger, *Viruses associated with human cancer.* Biochim Biophys Acta, 2008. 1782(3): p. 127-50.

145. Bakkalci, D., et al., *Risk factors for Epstein Barr virus-associated cancers: a systematic review, critical appraisal, and mapping of the epidemiological evidence.* J Glob Health, 2020. 10(1): p. 010405.

146. Ananthakrishnan, A., V. Gogineni, and K. Saeian, *Epidemiology of primary and secondary liver cancers.* Semin Intervent Radiol, 2006. 23(1): p. 47-63.

147. Byun, J.M., et al., *Persistent HPV-16 infection leads to recurrence of high-grade cervical intraepithelial neoplasia.* Medicine (Baltimore), 2018. 97(51): p. e13606.

148. Pal, A. and R. Kundu, *Human Papillomavirus E6 and E7: The Cervical Cancer Hallmarks and Targets for Therapy.* Front Microbiol, 2019. 10: p. 3116.

149. Cousins, E. and J. Nicholas, *Molecular biology of human herpesvirus 8: novel functions and virus-host interactions implicated in viral pathogenesis and replication.* Recent Results Cancer Res, 2014. 193: p. 227-68.

150. Hammerschmidt, W. and B. Sugden, *Epstein-Barr virus sustains Burkitt's lymphomas and Hodgkin's disease.* Trends Mol Med, 2004. 10(7): p. 331-6.

151. Chesson, H.W., et al., *The estimated lifetime probability of acquiring human papillomavirus in the United States.* Sex Transm Dis, 2014. 41(11): p. 660-4.

152. Amber, K., M.P. McLeod, and K. Nouri, *The Merkel cell polyomavirus and its involvement in Merkel cell carcinoma.* Dermatol Surg, 2013. 39(2): p. 232-8.

153. Jarrett, A., A. Armstrong, and E. Alexander, *Epidemiology of EBV and Hodgkin's lymphoma.* Annals of oncology, 1996. 7: p. S5-S10.

154. Vockerodt, M., et al., *Epstein-Barr virus and the origin of Hodgkin lymphoma.* Chin J Cancer, 2014. 33(12): p. 591-7.

155. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. 71(3): p. 209-249.

156. Abdulamir, A., et al., *The distinctive profile of risk factors of nasopharyngeal carcinoma in comparison with other head and neck cancer types.* BMC public health, 2008. 8(1): p. 1-16.

157. Adham, M., et al., *Nasopharyngeal carcinoma in Indonesia: epidemiology, incidence, signs, and symptoms at presentation.* Chinese journal of cancer, 2012. 31(4): p. 185.

158. Agelli, M. and L.X. Clegg, *Epidemiology of primary Merkel cell carcinoma in the United States.* Journal of the American Academy of Dermatology, 2003. 49(5): p. 832-841.

159. Ajila, V., et al., *Human papilloma virus associated squamous cell carcinoma of the head and neck.* Journal of sexually transmitted diseases, 2015. 2015: p. 791024.

160. Aka, P., et al., *Incidence and trends in Burkitt lymphoma in northern Tanzania from 2000 to 2009.* Pediatric blood & cancer, 2012. 59(7): p. 1234-1238.

161. Akhtar, S., K.K. Oza, and J. Wright, *Merkel cell carcinoma: report of 10 cases and review of the literature.* Journal of the American Academy of Dermatology, 2000. 43(5): p. 755-767.

162. Alipov, G., et al., *Epstein-Barr virus-associated gastric carcinoma in Kazakhstan.* World Journal of Gastroenterology, 2005. 11(1): p. 27.

163. Andres, C., et al., *Prevalence of MCPyV in Merkel cell carcinoma and non-MCC tumors.* Journal of cutaneous pathology, 2010. 37(1): p. 28-34.

164. Bassig, B.A., et al., *Subtype-specific incidence rates of lymphoid malignancies in Hong Kong compared to the United States, 2001-2010.* Cancer Epidemiology, 2016. 42: p. 15-23.

165. Bosch, F.X., et al., *Epidemiology of hepatocellular carcinoma.* Clinics in liver disease, 2005. 9(2): p. 191-211.

166. Carrascal, E., et al., *Epstein-Barr virus-associated gastric carcinoma in Cali, Colombia.* Oncology reports, 2003. 10(4): p. 1059-1062.

167. Chang, M.-H., et al., *Hepatitis B vaccination and hepatocellular carcinoma rates in boys and girls.* Jama, 2000. 284(23): p. 3040-3042.

168. Chang, M.S., et al., *Clinicopathologic characteristics of Epstein-Barr virus-incorporated gastric cancers in Korea.* Pathology-Research and Practice, 2001. 197(6): p. 395-400.

169. Chen, W., et al., *Esophageal cancer incidence and mortality in China, 2009.* Journal of thoracic disease, 2013. 5(1): p. 19.

170. Chong, J.M., et al., *Expression of CD44 variants in gastric carcinoma with or without Epstein-Barr virus.* International journal of cancer, 1997. 74(4): p. 450-454.

171. Claviez, A., et al., *Impact of latent Epstein-Barr virus infection on outcome in children and adolescents with Hodgkin's lymphoma.* Journal of clinical oncology, 2005. 23(18): p. 4048-4056.

172. Conte, S., et al., *Population-Based Study detailing cutaneous melanoma incidence and mortality trends in Canada.* Frontiers in medicine, 2022. 9: p. 830254.

173. Corvalan, A., et al., *Epstein-Barr virus in gastric carcinoma is associated with location in the cardia and with a diffuse histology: a study in one area of Chile.* International journal of cancer, 2001. 94(4): p. 527-530.

174. Czopek, J.P., et al., *EBV-positive gastric carcinomas in Poland.* Polish Journal of Pathology: Official Journal of the Polish Society of Pathologists, 2003. 54(2): p. 123-128.

175. Deo, S., et al., *Colorectal Cancers in Low-and Middle-Income Countries—Demographic Pattern and Clinical Profile of 970 Patients Treated at a Tertiary Care Cancer Center in India.* JCO Global Oncology, 2021. 7: p. 1110-1115.

176. Diepstra, A., et al., *Latent Epstein-Barr virus infection of tumor cells in classical Hodgkin's lymphoma predicts adverse outcome in older adult patients.* J Clin Oncol, 2009. 27(23): p. 3815-21.

177. Divaris, K., et al., *Oral health and risk for head and neck squamous cell carcinoma: the Carolina Head and Neck Cancer Study.* Cancer Causes & Control, 2010. 21(4): p. 567-575.

178. Enblad, G., et al., *Epstein-Barr virus distribution in Hodgkin's disease in an unselected Swedish population.* Acta Oncologica, 1999. 38(4): p. 425-429.

179. Fedder, M. and M.F. Gonzalez, *Nasopharyngeal carcinoma. Brief review.* The American journal of medicine, 1985. 79(3): p. 365-369.

180. Galetsky, S.A., et al., *Epstein-Barr-virus-associated gastric cancer in Russia.* International Journal of Cancer, 1997. 73(6): p. 786-789.

181. Gulley, M.L., et al., *Epstein-Barr virus infection is an early event in gastric carcinogenesis and is independent of bcl-2 expression and p53 accumulation.* Human pathology, 1996. 27(1): p. 20-27.

182. Hao, Z., et al., *The Epstein-Barr virus-associated gastric carcinoma in Southern and Northern China.* Oncology reports, 2002. 9(6): p. 1293-1298.

183. Harn, H.-J., et al., *Epstein-Barr virus-associated gastric adenocarcinoma in Taiwan.* Human pathology, 1995. 26(3): p. 267-271.

184. Herrera-Goepfert, R., et al., *Epstein-Barr virus-associated gastric carcinoma: Evidence of age-dependence among a Mexican population.* World journal of gastroenterology, 2005. 11(39): p. 6096.

185. Hjalgrim, H., J. Friborg, and M. Melbye, *The epidemiology of EBV and its association with malignant disease*, in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*, A. Arvin, et al., Editors. 2007, Cambridge University Press: Cambridge.

186. Hsu, J.L. and S.L. Glaser, *Epstein–Barr virus-associated malignancies: epidemiologic patterns and etiologic implications.* Critical reviews in oncology/hematology, 2000. 34(1): p. 27-53.

187. Iscovich, J., et al., *Classic Kaposi sarcoma: epidemiology and risk factors.* Cancer, 2000. 88(3): p. 500-517.

188. Jarrett, R., et al., *The Scotland and Newcastle epidemiological study of Hodgkin's disease: impact of histopathological review and EBV status on incidence estimates.* Journal of clinical pathology, 2003. 56(11): p. 811-816.

189. Johansson, S.L. and S.M. Cohen, *Epidemiology and etiology of bladder cancer.* Seminars in surgical oncology, 1997. 13(5): p. 291-298.

190. Kang, G.H., et al., *Epstein-barr virus-positive gastric carcinoma demonstrates frequent aberrant methylation of multiple genes and constitutes CpG island methylator phenotype-positive gastric carcinoma.* The American journal of pathology, 2002. 160(3): p. 787-794.

191. Karim, N. and G. Pallesen, *Epstein-Barr virus (EBV) and gastric carcinoma in Malaysian patients.* The Malaysian journal of pathology, 2003. 25(1): p. 45-47.

192. Kassem, A., et al., *Frequent detection of Merkel cell polyomavirus in human Merkel cell carcinomas and identification of a unique deletion in the VP1 gene.* Cancer research, 2008. 68(13): p. 5009-5013.

193. Keegan, T.H., et al., *Epstein-Barr virus as a marker of survival after Hodgkin's lymphoma: a population-based study.* Journal of Clinical Oncology, 2005. 23(30): p. 7604-7613.

194. Koriyama, C., et al., *Epstein-Barr virus-associated gastric carcinoma in Japanese Brazilians and non-Japanese Brazilians in Sao Paulo.* Japanese journal of cancer research, 2001. 92(9): p. 911-917.

195. Kume, T., et al., *Low rate of apoptosis and overexpression of bcl-2 in Epstein–Barr virus-associated gastric carcinoma.* Histopathology, 1999. 34(6): p. 502-509.

196. Lopes, L., et al., *Epstein-Barr virus infection and gastric carcinoma in São Paulo State, Brazil.* Brazilian Journal of Medical and Biological Research, 2004. 37: p. 1707-1712.

197. Lu, S.N., et al., *Secular trends and geographic variations of hepatitis B virus and hepatitis C virus-associated hepatocellular carcinoma in Taiwan.* International journal of cancer, 2006. 119(8): p. 1946-1952.

198. Mbulaiteye, S.M., et al., *Trimodal age-specific incidence patterns for Burkitt lymphoma in the United States, 1973–2005.* International journal of cancer, 2010. 126(7): p. 1732-1739.

199. McGlynn, K.A. and W.T. London, *Epidemiology and natural history of hepatocellular carcinoma.* Best practice & research Clinical gastroenterology, 2005. 19(1): p. 3-23.

200. McNeil, D.E., et al., *SEER update of incidence and trends in pediatric malignancies: acute lymphoblastic leukemia.* Medical and pediatric oncology, 2002. 39(6): p. 554-557.

201. Mimi, C.Y. and J.-M. Yuan, *Epidemiology of nasopharyngeal carcinoma.* Seminars in cancer biology, 2002. 12(6): p. 421-429.

202. Molica, S., *Sex differences in incidence and outcome of chronic lymphocytic leukemia patients.* Leukemia & lymphoma, 2006. 47(8): p. 1477-1480.

203. Moritani, S., et al., *Phenotypic characteristics of Epstein-Barr-virus-associated gastric carcinomas.* Journal of cancer research and clinical oncology, 1996. 122(12): p. 750-756.

204. Murphy, G., et al., *Meta-analysis shows that prevalence of Epstein–Barr virus-positive gastric cancer differs based on sex and anatomic location.* Gastroenterology, 2009. 137(3): p. 824-833.

205. Nogueira, C., et al., *Prevalence and characteristics of Epstein–Barr virus-associated gastric carcinomas in Portugal.* Infectious agents and cancer, 2017. 12(1): p. 1-8.

206. Ogwang, M.D., et al., *Incidence and geographic distribution of endemic Burkitt lymphoma in northern Uganda revisited.* International journal of cancer, 2008. 123(11): p. 2658-2663.

207. Ojima, H., et al., *Discrepancy between clinical and pathological lymph node evaluation in Epstein-Barr virus-associated gastric cancers.* Anticancer research, 1996. 16(5B): p. 3081-3084.

208. Pallagani, L., et al., *Epidemiology and clinicopathological profile of renal cell carcinoma: a review from tertiary care referral centre.* Journal of Kidney Cancer and VHL, 2021. 8(1): p. 1.

209. Qiu, K., et al., *Epstein-Barr virus in gastric carcinoma in Suzhou, China and Osaka, Japan: Association with clinico-pathologic factors and HLA-subtype.* International journal of cancer, 1997. 71(2): p. 155-158.

210. Ragin, C., F. Modugno, and S. Gollin, *The epidemiology and risk factors of head and neck cancer: a focus on human papillomavirus.* Journal of dental research, 2007. 86(2): p. 104-114.

211. Rahbari, R., L. Zhang, and E. Kebebew, *Thyroid cancer gender disparity.* Future Oncology, 2010. 6(11): p. 1771-1779.

212. Randi, G., S. Franceschi, and C. La Vecchia, *Gallbladder cancer worldwide: geographical distribution and risk factors.* International journal of cancer, 2006. 118(7): p. 1591-1602.

213. Rawla, P. and A. Barsouk, *Epidemiology of gastric cancer: global trends, risk factors and prevention.* Gastroenterology Review/Przegląd Gastroenterologiczny, 2019. 14(1): p. 26-38.

214. Rowlands, D., et al., *Epstein-Barr virus and carcinomas: rare association of the virus with gastric adenocarcinomas.* British journal of cancer, 1993. 68(5): p. 1014-1019.

215. Sakuma, K., et al., *Cancer risk to the gastric corpus in Japanese, its correlation with interleukin-1β gene polymorphism (+ 3953* T) and Epstein-Barr virus infection.* International journal of cancer, 2005. 115(1): p. 93-97.

216. Sellam, F., et al., *Delayed diagnosis of pancreatic cancer reported as more common in a population of North African young adults.* Journal of gastrointestinal oncology, 2015. 6(5): p. 505.

217. Shibata, D. and L. Weiss, *Epstein-Barr virus-associated gastric adenocarcinoma.* The American journal of pathology, 1992. 140(4): p. 769.

218. Shin, W.S., et al., *Epstein-Barr virus-associated gastric adenocarcinomas among Koreans.* American journal of clinical pathology, 1996. 105(2): p. 174-181.

219. Souza, E.M., et al., *Impact of Epstein–Barr virus in the clinical evolution of patients with classical Hodgkin's lymphoma in Brazil.* Hematological Oncology, 2010. 28(3): p. 137-141.

220. Takano, Y., et al., *The role of the Epstein-Barr virus in the oncogenesis of EBV (+) gastric carcinomas.* Virchows Archiv, 1999. 434(1): p. 17-22.

221. Tamási, L., et al., *Age and Gender Specific Lung Cancer Incidence and Mortality in Hungary: Trends from 2011 Through 2016.* Pathology and Oncology Research, 2021: p. 88.

222. Tamimi, A.F. and M. Juweid, *Epidemiology and outcome of glioblastoma.* Exon Publications, 2017: p. 143-153.

223. Tavakoli, A., et al., *Association between Epstein-Barr virus infection and gastric cancer: a systematic review and meta-analysis.* BMC cancer, 2020. 20(1): p. 1-14.

224. Tokunaga, M., et al., *Epstein-Barr virus in gastric carcinoma.* The American journal of pathology, 1993. 143(5): p. 1250.

225. van Beek, J., et al., *EBV-positive gastric adenocarcinomas: a distinct clinicopathologic entity with a low frequency of lymph node involvement.* Journal of Clinical Oncology, 2004. 22(4): p. 664-670.

226. Venook, A.P., et al., *The incidence and epidemiology of hepatocellular carcinoma: a global and regional perspective.* The oncologist, 2010. 15(S4): p. 5-13.

227. Villano, J., et al., *Age, gender, and racial differences in incidence and survival in primary CNS lymphoma.* British journal of cancer, 2011. 105(9): p. 1414-1418.

228. Wang, X.m., et al., *Clinical analysis of 1629 newly diagnosed malignant lymphomas in current residents of Sichuan province, China.* Hematological oncology, 2016. 34(4): p. 193-199.

229. Wang, Y., et al., *Quantitative methylation analysis reveals gender and age differences in p16 INK 4a hypermethylation in hepatitis B virus-related hepatocellular carcinoma.* Liver International, 2012. 32(3): p. 420-428.

230. Wei, K.-R., et al., *Nasopharyngeal carcinoma incidence and mortality in China in 2010.* Chinese journal of cancer, 2014. 33(8): p. 381.

231. Wu, M.S., et al., *Epstein–Barr virus—associated gastric carcinomas: relation to H. pylori infection and genetic alterations.* Gastroenterology, 2000. 118(6): p. 1031-1038.

232. Wu, S., et al., *Basal-cell carcinoma incidence and associated risk factors in US women and men.* American journal of epidemiology, 2013. 178(6): p. 890-897.

233. Yanai, H., et al., *Endoscopic and pathologic features of Epstein-Barr virus–associated gastric carcinoma.* Gastrointestinal endoscopy, 1997. 45(3): p. 236-242.

234. Yoshiwara, E., et al., *Epstein-Barr virus-associated gastric carcinoma in Lima, Peru.* J Exp Clin Cancer Res, 2005. 24(1): p. 49-54.

235. Zhou, L., et al., *Global, regional, and national burden of Hodgkin lymphoma from 1990 to 2017: estimates from the 2017 Global Burden of Disease study.* Journal of hematology & oncology, 2019. 12(1): p. 1-13.

236. Zhu, Z.-Z., et al., *Sex-related differences in DNA copy number alterations in hepatitis B virus-associated hepatocellular carcinoma.* Asian Pacific Journal of Cancer Prevention, 2012. 13(1): p. 225-229.

237. Rismiller, K. and T.J. Knackstedt, *Aggressive Digital Papillary Adenocarcinoma: Population-Based Analysis of Incidence, Demographics, Treatment, and Outcomes.* Dermatol Surg, 2018. 44(7): p. 911-917.

238. Li, X., et al., *HLA associations with nasopharyngeal carcinoma.* Current molecular medicine, 2009. 9(6): p. 751-765.

239. Huang, X., et al., *HLA-A* 02: 07 is a protective allele for EBV negative and a susceptibility allele for EBV positive classical Hodgkin lymphoma in China.* PLoS One, 2012. 7(2): p. e31865.

240. Schottenfeld, D. and J. Beebe-Dimmer, *The cancer burden attributable to biologic agents.* Annals of Epidemiology, 2015. 25(3): p. 183-187.

241. Jemal, A., et al., *Global Patterns of Cancer Incidence and Mortality Rates and TrendsGlobal Patterns of Cancer.* Cancer epidemiology, biomarkers & prevention, 2010. 19(8): p. 1893-1907.

242. The Cancer Genome Atlas Research Network, *Comprehensive molecular characterization of gastric adenocarcinoma.* Nature, 2014. 513(7517): p. 202-9.

243. The Cancer Genome Atlas Research Network, *Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma.* Cell, 2017. 169(7): p. 1327-1341 e23.

244. The Cancer Genome Atlas Research Network, *Integrated genomic and molecular characterization of cervical cancer.* Nature, 2017. 543(7645): p. 378-384.

245. The Cancer Genome Atlas Network, *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. 517(7536): p. 576-82.

246. Ramis-Zaldivar, J.E., et al., *MAPK and JAK-STAT pathways dysregulation in plasmablastic lymphoma.* Haematologica, 2021. 106(10): p. 2682-2693.

247. Gandhi, M.K., et al., *EBV-associated primary CNS lymphoma occurring after immunosuppression is a distinct immunobiological entity.* Blood, 2021. 137(11): p. 1468-1477.

248. Starrett, G.J., et al., *Clinical and molecular characterization of virus-positive and virus-negative Merkel cell carcinoma.* Genome Med, 2020. 12(1): p. 30.

249. Grande, B.M., et al., *Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma.* Blood, 2019. 133(12): p. 1313-1324.

250. Alexandrov, L.B., et al., *The repertoire of mutational signatures in human cancer.* Nature, 2020. 578(7793): p. 94-101.

251. Henderson, S., et al., *APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development.* Cell reports, 2014. 7(6): p. 1833-1841.

252. Warren, C.J., et al., *Roles of APOBEC3A and APOBEC3B in human papillomavirus infection and disease progression.* Viruses, 2017. 9(8): p. 233.

253. Chu, Y.-J., et al., *Aflatoxin B1 exposure increases the risk of hepatocellular carcinoma associated with hepatitis C virus infection or alcohol consumption.* European Journal of Cancer, 2018. 94: p. 37-46.

254. Chu, Y.J., et al., *Aflatoxin B1 exposure increases the risk of cirrhosis and hepatocellular carcinoma in chronic hepatitis B virus carriers.* International journal of cancer, 2017. 141(4): p. 711-720.

255. Rocak, S. and P. Linder, *DEAD-box proteins: the driving forces behind RNA metabolism.* Nat Rev Mol Cell Biol, 2004. 5(3): p. 232-41.

256. Dufva, O., et al., *Aggressive natural killer-cell leukemia mutational landscape and drug profiling highlight JAK-STAT signaling as therapeutic target.* Nat Commun, 2018. 9(1): p. 1567.

257. Kataoka, K., et al., *Integrated molecular analysis of adult T cell leukemia/lymphoma.* Nat Genet, 2015. 47(11): p. 1304-15.

258. Xiong, J., et al., *Genomic and Transcriptomic Characterization of Natural Killer T Cell Lymphoma.* Cancer Cell, 2020. 37(3): p. 403-419 e6.

259. Zhang, L., et al., *Genomic Analysis of Nasopharyngeal Carcinoma Reveals TME-Based Subtypes.* Mol Cancer Res, 2017. 15(12): p. 1722-1732.

260. Mo, J., et al., *DDX3X: structure, physiologic functions and cancer.* Mol Cancer, 2021. 20(1): p. 38.

261. Gong, C., et al., *Sequential inverse dysregulation of the RNA helicases DDX3X and DDX3Y facilitates MYC-driven lymphomagenesis.* Mol Cell, 2021. 81(19): p. 4059-4075 e11.

262. Thomas, N., et al., *Genetic Subgroups Inform on Pathobiology in Adult and Pediatric Burkitt Lymphoma.* Blood, 2022. blood.2022016534.

263. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. 45(6): p. 580-5.

264. Formica, V., et al., *A systematic review and meta-analysis of PD-1/PD-L1 inhibitors in specific patient subgroups with advanced gastro-oesophageal junction and gastric adenocarcinoma.* Crit Rev Oncol Hematol, 2021. 157: p. 103173.

265. De Meulenaere, A., et al., *Turning the tide: Clinical utility of PD-L1 expression in squamous cell carcinoma of the head and neck.* Oral Oncol, 2017. 70: p. 34-42.

266. Lipson, E.J., et al., *PD-L1 expression in the Merkel cell carcinoma microenvironment: association with inflammation, Merkel cell polyomavirus and overall survival.* Cancer Immunol Res, 2013. 1(1): p. 54-63.

267. Derks, S., et al., *Abundant PD-L1 expression in Epstein-Barr Virus-infected gastric cancers.* Oncotarget, 2016. 7(22): p. 32925-32.

268. Yang, W.F., et al., *The prognostic role of PD-L1 expression for survival in head and neck squamous cell carcinoma: A systematic review and meta-analysis.* Oral Oncol, 2018. 86: p. 81-90.

269. Li, B., et al., *Anti-PD-1/PD-L1 Blockade Immunotherapy Employed in Treating Hepatitis B Virus Infection-Related Advanced Hepatocellular Carcinoma: A Literature Review.* Front Immunol, 2020. 11: p. 1037.

270. Blumberg, B.S., *The curiosities of hepatitis B virus: prevention, sex ratio, and demography.* Proc Am Thorac Soc, 2006. 3(1): p. 14-20.

271. Fish, E.N., *The X-files in immunity: sex-based differences predispose immune responses.* Nat Rev Immunol, 2008. 8(9): p. 737-44.

272. Klein, S.L. and K.L. Flanagan, *Sex differences in immune responses.* Nat Rev Immunol, 2016. 16(10): p. 626-38.

273. Zhang, B.L., et al., *Somatic mutation profiling of liver and biliary cancer by targeted next generation sequencing.* Oncol Lett, 2018. 16(5): p. 6003-6012.

274. Zhao, L.H., et al., *Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma.* Nat Commun, 2016. 7: p. 12992.

275. Tornesello, M.L., et al., *Mutations in TP53, CTNNB1 and PIK3CA genes in hepatocellular carcinoma associated with hepatitis B and hepatitis C virus infections.* Genomics, 2013. 102(2): p. 74-83.

276. Riva, V. and G. Maga, *From the magic bullet to the magic target: exploiting the diverse roles of DDX3X in viral infections and tumorigenesis.* Future Med Chem, 2019. 11(11): p. 1357-1381.

277. He, Y., et al., *A double-edged function of DDX3, as an oncogene or tumor suppressor, in cancer progression (Review).* Oncol Rep, 2018. 39(3): p. 883-892.

278. Chatterjee, K., et al., *The interplay between Epstein-Bar virus (EBV) with the p53 and its homologs during EBV associated malignancies.* Heliyon, 2019. 5(11): p. e02624.

279. Miliotis, C.N. and F.J. Slack, *Multi-layered control of PD-L1 expression in Epstein-Barr virus-associated gastric cancer.* J Cancer Metastasis Treat, 2020. 6(13).

280. Ukpo, O.C., W.L. Thorstad, and J.S. Lewis, Jr., *B7-H1 expression model for immune evasion in human papillomavirus-related oropharyngeal squamous cell carcinoma.* Head Neck Pathol, 2013. 7(2): p. 113-21.

281. Lyford-Pike, S., et al., *Evidence for a role of the PD-1:PD-L1 pathway in immune resistance of HPV-associated head and neck squamous cell carcinoma.* Cancer Res, 2013. 73(6): p. 1733-41.

282. Kim, H.S., et al., *Association Between PD-L1 and HPV Status and the Prognostic Value of PD-L1 in Oropharyngeal Squamous Cell Carcinoma.* Cancer Res Treat, 2016. 48(2): p. 527-36.

283. Badoual, C., et al., *PD-1-expressing tumor-infiltrating T cells are a favorable prognostic biomarker in HPV-associated head and neck cancer.* Cancer Res, 2013. 73(1): p. 128-38.

284. Oliva, M., et al., *Immune biomarkers of response to immune-checkpoint inhibitors in head and neck squamous cell carcinoma.* Ann Oncol, 2019. 30(1): p. 57-67.

285. Beerenwinkel, N., et al., *Genetic progression and the waiting time to cancer.* PLoS Comput Biol, 2007. 3(11): p. e225.

286. Tso, F.Y., et al., *RNA-Seq of Kaposi's sarcoma reveals alterations in glucose and lipid metabolism.* PLoS Pathog, 2018. 14(1): p. e1006844.

287. Orenbuch, R., et al., *arcasHLA: high-resolution HLA typing from RNAseq.* Bioinformatics, 2020. 36(1): p. 33-40.

288. Thorsson, V., et al., *The Immune Landscape of Cancer*. Immunity, 2018. 48(4): p. 812-830 e14.

289. Swerdlow, S.H., et al., *WHO classification of tumours of haematopoietic and lymphoid tissues*. Vol. 2. 2008: International agency for research on cancer Lyon.

290. Carbone, A., et al., *Diagnosis and management of lymphomas and other cancers in HIV-infected patients.* Nature reviews Clinical oncology, 2014. 11(4): p. 223-238.

291. Noy, A., et al., *Plasmablastic lymphoma is treatable in the HAART era. A 10 year retrospective by the AIDS Malignancy Consortium.* Leukemia & lymphoma, 2016. 57(7): p. 1731-1734.

292. Pather, S., et al., *Large cell lymphoma: correlation of HIV status and prognosis with differentiation profiles assessed by immunophenotyping.* Pathology & Oncology Research, 2013. 19(4): p. 695-705.

293. Delecluse, H., et al., *Plasmablastic lymphomas of the oral cavity: a new entity associated with the human immunodeficiency virus infection.* Blood, 1997. 89(4): p. 1413-1420.

294. Valera, A., et al., *IG/MYC rearrangements are the main cytogenetic alteration in plasmablastic lymphomas.* The American journal of surgical pathology, 2010. 34(11): p. 1686.

295. Tchernonog, E., et al., *Clinical characteristics and prognostic factors of plasmablastic lymphoma patients: analysis of 135 patients from the LYSA group.* Annals of Oncology, 2017. 28(4): p. 843-848.

296. Teruya-Feldstein, J., et al., *CD20-negative large-cell lymphoma with plasmablastic features: a clinically heterogenous spectrum in both HIV-positive and-negative patients.* Annals of Oncology, 2004. 15(11): p. 1673-1679.

297. Chang, C.-C., et al., *Genomic profiling of plasmablastic lymphoma using array comparative genomic hybridization (aCGH): revealing significant overlapping genomic lesions with diffuse large B-cell lymphoma.* Journal of hematology & oncology, 2009. 2(1): p. 1-6.

298. Chapman, J., et al., *Gene expression analysis of plasmablastic lymphoma identifies downregulation of B-cell receptor signaling and additional unique transcriptional programs.* Leukemia, 2015. 29(11): p. 2270-2273.

299. Montes-Moreno, S., et al., *Plasmablastic lymphoma phenotype is determined by genetic alterations in MYC and PRDM1*. Modern Pathology, 2017. 30(1): p. 85-94.

300. Hillmer, E.J., et al., *STAT3 signaling in immunity.* Cytokine & growth factor reviews, 2016. 31: p. 1-15.

301. Crescenzo, R., et al., *Convergent mutations and kinase fusions lead to oncogenic STAT3 activation in anaplastic large cell lymphoma.* Cancer cell, 2015. 27(4): p. 516-532.

302. Song, T.L., et al., *Oncogenic activation of the STAT3 pathway drives PD-L1 expression in natural killer/T-cell lymphoma.* Blood, 2018. 132(11): p. 1146-1158.

303. Nicolae, A., et al., *Mutations in the JAK/STAT and RAS signaling pathways are common in intestinal T-cell lymphomas.* Leukemia, 2016. 30(11): p. 2245-2247.

304. Prior, I.A., P.D. Lewis, and C. Mattos, *A comprehensive survey of Ras mutations in cancer.* Cancer research, 2012. 72(10): p. 2457-2467.

305. Pasqualucci, L., et al., *Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas.* Nature, 2001. 412(6844): p. 341-346.

306. Taddesse-Heath, L., et al., *Plasmablastic lymphoma with MYC translocation: evidence for a common pathway in the generation of plasmablastic features.* Modern Pathology, 2010. 23(7): p. 991-999.

307. Boy, S.C., et al., *Dominant genetic aberrations and coexistent EBV infection in HIV-related oral plasmablastic lymphomas.* Oral oncology, 2011. 47(9): p. 883-887.

308. Miles, D.M., et al., *High levels of histones promote whole-genome-duplications and trigger a Swe1WEE1-dependent phosphorylation of Cdc28CDK1.* Elife, 2018. 7: p. e35337.

309. Wenzel, S., et al., *MCL1 is deregulated in subgroups of diffuse large B-cell lymphoma.* Leukemia, 2013. 27(6): p. 1381-1390.

310. Chen, C., et al., *The biology and role of CD44 in cancer progression: therapeutic implications.* Journal of hematology & oncology, 2018. 11(1): p. 1-23.

311. Landau, D.A., et al., *Mutations driving CLL and their evolution in progression and relapse.* Nature, 2015. 526(7574): p. 525-530.

312. Lohr, J.G., et al., *Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy.* Cancer cell, 2014. 25(1): p. 91-101.

313. Pasqualucci, L. and R. Dalla-Favera, *Genetics of diffuse large B-cell lymphoma.* Blood, 2018. 131(21): p. 2307-2319.

314. Brescia, P., et al., *MEF2B instructs germinal center development and acts as an oncogene in B cell lymphomagenesis.* Cancer cell, 2018. 34(3): p. 453-465. e9.

315. Ramsay, A.J., et al., *Next-generation sequencing reveals the secrets of the chronic lymphocytic leukemia genome.* Clinical and Translational Oncology, 2013. 15(1): p. 3-8.

316. Chapman, M.A., et al., *Initial genome sequencing and analysis of multiple myeloma.* Nature, 2011. 471(7339): p. 467-472.

317. Gravelle, P., et al., *EBV infection determines the immune hallmarks of plasmablastic lymphoma.* Oncoimmunology, 2018. 7(10): p. e1486950.

318. Shannon-Lowe, C. and A. Rickinson, *The global landscape of EBV-associated tumors.* Frontiers in oncology, 2019. 9: p. 713.

319. Bencun, M., et al., *Translational profiling of B cells infected with the Epstein-Barr virus reveals 5′ leader ribosome recruitment through upstream open reading frames.* Nucleic acids research, 2018. 46(6): p. 2802-2819.

320. Hammerschmidt, W. and B. Sugden, *Replication of Epstein–Barr Viral DNA.* Cold Spring Harbor perspectives in biology, 2013. 5(1): p. a013029.

321. Sugimoto, A., et al., *Different distributions of Epstein-Barr virus early and late gene transcripts within viral replication compartments.* Journal of virology, 2013. 87(12): p. 6693-6699.

322. Yu, H., et al., *Revisiting STAT3 signalling in cancer: new and unexpected biological functions.* Nature reviews cancer, 2014. 14(11): p. 736-746.

323. Garcia-Reyero, J., et al., *Genetic lesions in MYC and STAT3 drive oncogenic transcription factor overexpression in plasmablastic lymphoma.* Haematologica, 2021. 106(4): p. 1120.

324. Cousins, E., et al., *Human herpesvirus 8 viral interleukin-6 signaling through gp130 promotes virus replication in primary effusion lymphoma and endothelial cells.* Journal of virology, 2014. 88(20): p. 12167-12172.

325. Chiarle, R., et al., *Stat3 is required for ALK-mediated lymphomagenesis and provides a possible therapeutic target.* Nature medicine, 2005. 11(6): p. 623-629.

326. Valera, A., et al., *ALK-positive large B-cell lymphomas express a terminal B-cell differentiation program and activated STAT3 but lack MYC rearrangements.* Modern Pathology, 2013. 26(10): p. 1329-1337.

327. Steidl, C. and R.D. Gascoyne, *The molecular pathogenesis of primary mediastinal large B-cell lymphoma.* Blood, 2011. 118(10): p. 2659-2669.

328. Joos, S., et al., *Genomic imbalances including amplification of the tyrosine kinase gene JAK2 in CD30+ Hodgkin cells.* Cancer research, 2000. 60(3): p. 549-552.

329. Ritz, O., et al., *Recurrent mutations of the STAT6 DNA binding domain in primary mediastinal B-cell lymphoma.* Blood, 2009. 114(6): p. 1236-1242.

330. Johnson, D.E., R.A. O'Keefe, and J.R. Grandis, *Targeting the IL-6/JAK/STAT3 signalling axis in cancer.* Nature reviews Clinical oncology, 2018. 15(4): p. 234-248.

331. Redondo-Muñoz, J., A. García-Pardo, and J. Teixidó, *Molecular players in hematologic tumor cell trafficking.* Frontiers in immunology, 2019. 10: p. 156.

332. Tzankov, A., et al., *Prognostic significance of CD44 expression in diffuse large B cell lymphoma of activated and germinal centre B cell-like types: a tissue microarray analysis of 90 cases.* Journal of clinical pathology, 2003. 56(10): p. 747-752.

333. Zhong, Y., et al., *CD44-targeted vesicles encapsulating granzyme B as artificial killer cells for potent inhibition of human multiple myeloma in mice.* Journal of Controlled Release, 2020. 320: p. 421-430.

334. Koganti, S., et al., *Cellular STAT3 functions via PCBP2 to restrain Epstein-Barr Virus lytic activation in B lymphocytes.* Journal of virology, 2015. 89(9): p. 5002-5011.

335. Meer, S., et al., *Extraoral plasmablastic lymphomas in a high human immunodeficiency virus endemic area.* Histopathology, 2020. 76(2): p. 212-221.

336. Villela, D., et al., *Efficient detection of chromosome imbalances and single nucleotide variants using targeted sequencing in the clinical setting.* European journal of medical genetics, 2017. 60(12): p. 667-674.

337. Østrup, O., et al., *Detection of copy number alterations in cell-free tumor DNA from plasma.* BBA clinical, 2017. 7: p. 120-126.

338. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nature biotechnology, 2011. 29(7): p. 644-652.

339. Arvey, A., et al., *An atlas of the Epstein-Barr virus transcriptome and epigenome reveals host-virus regulatory interactions.* Cell host & microbe, 2012. 12(2): p. 233-245.

340. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. 30(7): p. 923-930.

Appendix A

Genomic Characterization of HIV-Associated Plasmablastic Lymphoma Identifies Pervasive Mutations in the JAK–STAT Pathway[4]

Blood Cancer Discov 2020;1:112–25

A.1 Introduction

Plasmablastic lymphoma (PBL) is a highly aggressive lymphoma of preterminally differentiated B cells, which predominantly occurs in patients with human immunodeficiency virus (HIV)-related or iatrogenic immunodeficiency [289]. As an AIDS-defining illness with a dismal prognosis, PBL is a particularly compelling problem in the Sub-Saharan African region, which accounts for approximately 54% of the estimated 37.9 million people living with HIV.

Despite the national rollout offering combination anti-retroviral therapy (cART) in South Africa since 2004, the prevalence of HIV-associated mature B-cell lymphomas increases yearly [290]. The burden of HIV-associated lymphomas on the country's encumbered health are further compounded by the younger presentation age of HIV-infected patients, the challenging classification of these lymphomas, and an exceptionally aggressive clinical disease course that, although curable in a significant fraction of patients [291], often results in fatalities [292].

Previously being classified as diffuse large B-cell lymphoma (DLBCL), PBL was later recognized as a distinct entity [293] with well-defined histopathologic features including large plasmablastic or immunoblastic cell morphology, loss of B-cell lineage markers (CD20, PAX5), expression of plasmacytic differentiation markers (CD38, CD138, IRF4/MUM1, BLIMP1), a high

[4] Material in this chapter is published wholly in [101] by Zhaoqi Liu, Ioan Filip, Karen Gomez, Dewaldt Engelbrecht, Shabnum Meer, Pooja N. Lalloo, Pareen Patel, Yvonne Perner, Junfei Zhao, Jiguang Wang, Laura Pasqualucci, Raul Rabadan, and Pascale Willem.

proliferation index, and frequent infection by the Epstein-Barr virus (EBV; ref. [56]). While EBV infection and activating MYC translocations have been reported as the major features of these tumors in a number of studies [294], the molecular pathology of PBL remains elusive in both HIV-positive and -negative individuals. Numerous case studies and some cohorts have been well described and reviewed [56, 294-296]. Only two studies have partially explored genetic aberrations in this disease: a comparative genomic hybridization study showed a pattern of segmental gains in PBL that more closely resembled DLBCL than plasma cell myeloma [297]. The second study compared the transcriptional profiles of 15 PBL cases to DLBCL and extraosseous plasmacytoma [298], and showed that B-cell receptor signaling genes including CD79A/B, BLK, LYN, and SWAP70 among others, were significantly downregulated in PBL compared with DLBCL; in contrast, targets of MYB and the oncogene MYC, as well as genes reflecting the known plasmacytic immunophenotype of PBL, were overexpressed. More recently, MYC and PRDM1 were investigated for mutations, structural rearrangements, and copy number gains in 36 PBL cases [299]. PRDM1 mutations were found in 8 of 16 cases, frequently in association with MYC overexpression, suggesting a coordinate role in the pathogenesis of the disease. While these studies have shed light on this rare disorder, a systematic characterization of protein-changing alterations in PBL has not been performed.

The elucidation of genes and pathways that drive the initiation and maintenance of PBL is essential to better understand the biology of this cancer and, critically, to implement improved biomarkers and more effective treatment options. Here, we combined whole-exome and transcriptome sequencing followed by targeted resequencing of 110 HIV-associated PBL cases to elucidate the mutational, transcriptional, and copy number landscape of this disease. We show that mutations in various components of the JAK–STAT and MAPK–ERK pathway pervade this

lymphoma, revealing this signaling cascade as a central oncogenic driver of the disease and a candidate for targeted therapy.

A.2 Results

The Landscape of Somatic Mutations in PBL

To determine the mutational landscape of PBL, we performed whole-exome sequencing (WES) in a discovery panel of paired tumor and normal DNA collected from 15 HIV-positive patients, followed by deep targeted sequencing of the top 34 candidate genes in 95 additional cases (Supplementary Table S1 and S2; Methods). Somatic mutations were identified from WES data using SAVI2 [122], an empirical Bayesian method. Overall, 2,149 nonsynonymous somatic variants were found in the 15 discovery cases, with a median of 45 per case and a total of 1,528 affected genes, of which 1,461 were mutated in the tumor-dominant clone (>15% cancer-specific allele frequency; Supplementary Table S3). Among the 15 WES cases, one was hypermutated (case PJ030), showing 845 sequence variants that were confirmed by RNA-sequencing (RNA-seq; 92% with a read depth ≥ 10; Supplementary Fig. S1A–S1C). Candidate genes were then selected for an extension screen based on the following criteria: (i) mutated in at least 2 discovery cases, (ii) expressed in normal and/or malignant B cells, (iii) known as a cancer driver gene, and/or (iv) with an established role in B-cell differentiation (Supplementary Tables S4 and S5; Methods). The 34 selected genes were all annotated in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. The mean depth of coverage for WES was 54.0x. The mean depth of coverage for the targeted DNA sequencing was 121.7x and, on average, 99.7% of the target sequences were covered by at least 50 reads.

We found genes in the JAK–STAT, MAPK–ERK, and Notch signaling pathways were commonly mutated in our PBL cohort (Fig.1A–E). The most frequent genetic lesions affected the

JAK–STAT signaling pathway with, in total, 62% of cases (68/110) harboring a mutation in at least one of five genes (STAT3, JAK1, SOCS1, JAK2, and PIM1, a direct transcriptional target of STAT3/5 that functions as part of a negative feedback loop; Fig.1A). Among these, STAT3 was the predominant target of mutations, with 46 of 110 cases (42%) harboring clonal events (Supplementary Fig. S2A and S2B). With three exceptions, the identified variants were heterozygous missense mutations clustering in exons 19 to 22, encoding part of the SH2 domain that is required for STAT3 molecular activation via receptor association and tyrosine phosphodimer formation [300]. In particular, the majority of the mutations resulted in the amino acid changes Y640F (n = 11), D661V (n = 9), S614R (n = 5), and E616G/K (n = 4;Fig. 4;Fig.1B), which have been categorized as gain-of-function or likely gain-of-function mutations based on experimental validation (https://oncokb.org/gene/STAT3), and overlap with the STAT3 mutation pattern described in other aggressive B-cell lymphomas and T-cell and natural killer (NK) cell malignancies [301, 302]. Three additional cases showed a >75% variant allele frequency in the absence of copy number changes, consistent with a copy neutral loss of heterozygosis. Sanger-based resequencing of the involved STAT3 exons, performed in 23 cases, validated all computationally identified mutations on these samples (Fig.1F Supplementary Table S6).

In addition to STAT3, 15/110 PBL cases harbored heterozygous missense mutations of JAK1, encoding for a tyrosine kinase that phosphorylates STAT proteins. These events did not involve the canonical hotspot seen in many other cancers (loss-of-function K860Nfs), but occurred at a highly conserved amino acid position in the JH1 kinase domain, G1097D/V (Fig.1C). Mutations at the JAK1 G1097 codon were previously reported in intestinal T-cell lymphomas [303] and anaplastic large cell lymphoma [301], and the G1097D substitution has been documented as a gain-of-function event that triggers aberrant phosphorylation of STAT3, leading

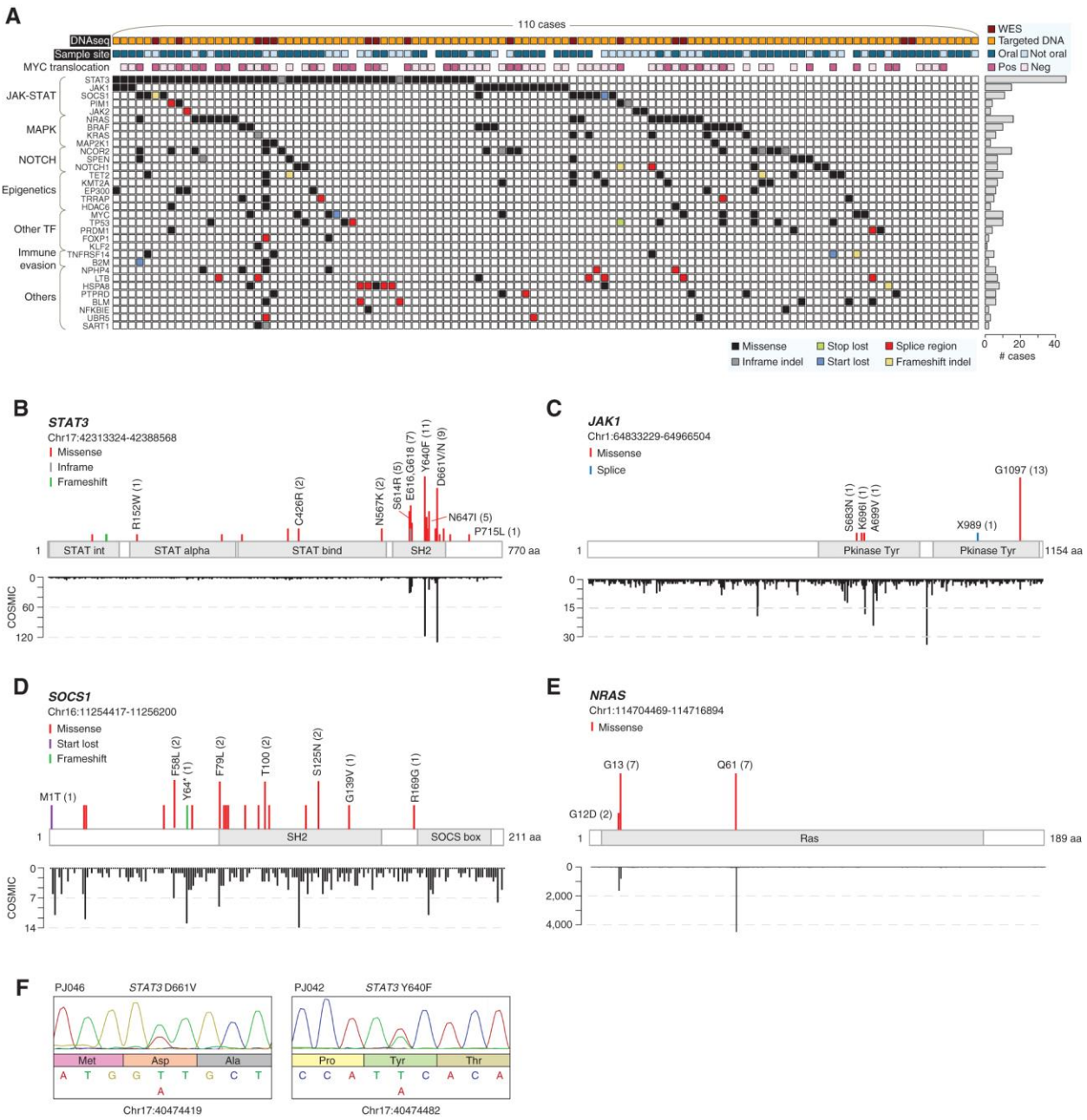


Figure 1: The landscape of putative driver gene mutations in PBL. A, Sample information, *MYC* translocation, and somatic mutation information are shown for 110 cases of PBL samples. The heatmap represents individual mutations in each sample, color-coded by type of mutation. B–E, Individual gene mutation maps for frequently mutated genes, showing mutation subtype, position, and evidence of mutational hotspots, based on COSMIC database information. *Y*-axis counts at the bottom of the maps reflect the number of identified mutations in the COSMIC database. F, Sanger validation of single nucleotide variants (SNV) in *STAT3* mutants.

to constitutive activation of the JAK–STAT signaling pathway [301]. Of note, missense mutations in STAT3 and JAK1 were found to be mutually exclusive (P = 0.04, Fisher exact test), suggesting a convergent role in activating this cascade. In contrast to STAT3 and JAK1, mutations in SOCS1 (n = 11/110 cases) were more widely distributed, consistent with its previously established role as a target of AID-mediated aberrant somatic hypermutation (Fig.1D). Although the consequences of most SOCS1 amino acid changes will require functional validation, the presence of a start-loss mutation in one case and a nonsense mutation in a second sample is expected to result in loss-of-function and inactivation of this negative JAK-STAT regulator, consistent with a tumor suppressor role.

The second most commonly mutated program was the MAPK–ERK signaling pathway, affected in 28% of cases by mutations in the RAS gene family members NRAS (14%) and KRAS (9%), as well as in BRAF (5.5%) and MAP2K1 (3%). Nearly all (92.5%) RAS mutations were found at known functional hotspots that have been shown to affect the intrinsic RAS GTPase activity, namely G12, G13, and Q61 (ref. [304]; Fig.1E). BRAF, an integral component of the MAPK signaling cascade, was found mutated in 6 cases, including 3 harboring mutations at or around the well-characterized V600 residue (namely: V600E, K601N, and T599TT) and 3 showing mutations at other common hotspots within the protein kinase domain (G464E, V471F, and M689V). The Valine at position 600 normally stabilizes the interaction between the BRAF glycine-rich loop and the activation segments, and its glutamine substitution confers an over 500-fold increase in activity, leading to constitutive activation of the MEK/ERK signaling cascade in the absence of extracellular stimuli. In line with their predicted gain of function, the mutant alleles in both RAS family members and BRAF were observed in heterozygosis (Supplementary Tables S3 and S5) and were actively expressed (median RPKM value of KRAS: 13.253, NRAS: 35.601,

and BRAF: 9.906).

In addition, 24% of PBL cases carried mutations in genes implicated in the Notch signaling pathway, including those encoding for NOTCH1, its negative regulator SPEN, and the Notch pathway corepressor NCOR2. In particular, one sample displayed a frameshift variant in the NOTCH1 gene that is predicted to generate a truncated protein lacking the C-terminal PEST domain, and thus endowed with increased protein half-life, while 6 additional cases showed amino acid changes within the EGF repeats, the juxtamembrane heterodimerization domain (N-terminal portion), and the C-terminal PEST domain. SPEN mutations include monoallelic missense substitutions that were distributed along the protein coding exons with no particular clustering, and their functional effect remains to be determined.

Missense mutations, frequently multiple within the same allele, were also found in the MYC gene (10/110, 9%), with 4 additional cases displaying silent mutations that possibly reflect the aberrant activity of the physiologic somatic hypermutation mechanism [305] as well as the recruitment of the AID mutator enzyme by the juxtaposed immunoglobulin enhancer in cases harboring chromosomal MYC translocations. MYC rearrangements with the immunoglobulin genes are the most common cytogenetic feature in PBL, present in about half of reported cases [294, 299, 306], and were identified in 36% (31/76) of our samples using FISH (Fig.1A). The functional consequences of MYC mutations will have to be experimentally tested; however, transcriptomic analysis revealed significantly higher expression of the MYC RNA in cases positive for the MYC translocation, consistent with MYC oncogenic activation (Supplementary Fig. S3A and S3B). The high proportion of cases showing evidence of MYC dysregulation reinforces the notion that MYC is a critical contributor to this aggressive cancer.

Aside from the genes mentioned above, we observed recurrent mutations, including

truncating loss-of-function events, in TET2 (10/110, 9%), TP53 (10/110, 9%), and NPHP4 (6/110, 5%). Finally, genes encoding epigenetic modifiers, transcription factors implicated in B-cell activation (FOXP1), positioning (KLF2), and terminal differentiation (PRDM1), and receptor molecules involved in tumor immune surveillance (e.g., B2M, TNFRSF14) were mutated to a lesser extent (range: 1%–5%; Fig.11A).

Although the relatively small number of cases prevents robust statistical analyses, we did not find any significant difference in the rate of mutated genes between samples collected from the oral cavity and from other locations (Pearson correlation coefficient = 0.822, Supplementary Fig. S2B), suggesting a genetic homogeneity of the disease regardless of the tissue of origin.

Collectively, the data presented above uncover a pervasive role for mutations affecting the JAK–STAT and MAPK–ERK pathways in the genetic landscape of PBL, which may contribute to the pathogenesis of this lymphoma by enforcing constitutive signaling activation, in concert with dysregulated MYC activity.

Somatic Copy Number Changes

To identify recurrent copy number alterations (CNA) associated with PBL, we applied the SNP-FASST2 algorithm to WES data from the 15 discovery cases (Supplementary Fig. S4A), followed by validation in an independent cohort of 31 additional PBL samples that had been processed with Affymetrix OncoScan microarrays (ref. [307]; Supplementary Table S7). Comparison of CNA calls obtained in parallel from the WES and OncoScan approach in a subset of 9 samples confirmed the data were highly consistent with each other (Supplementary Fig. S4B), supporting the robustness of the analysis.

Frequent copy number gains involved large chromosomal regions (>10 Mb) including chromosome 1q (20/46 cases, 43%) and the whole or most of chromosome 7 (13/46 cases, 28%).

When applying the GISTIC 2.0 algorithm to the combined cohort of 46 cases, we uncovered regions of highly recurrent amplification including 6p22.2 (the most significant), 6p22.1 and 1q21.3, all of which encompass histone gene clusters (Fig.2A). In particular, the chromosome 6p gain covered 36 genes that encode canonical histones and has been previously reported as a common alteration in a variety of cancers, where histone gains have been linked to genetic instability [308]. The significant region on chromosome 1q21.3 also included the IL6R gene and the antiapoptotic MCL1 gene, which showed increased gene expression (Fig.2C) analogous to what has been described in 26% of activated B-cell like (ABC) DLBCLs [309]. Although further studies will be needed to determine the functional impact of these alterations in PBL, chromosome 1q gains have been associated with unfavorable prognosis in multiple myeloma, suggesting a role in the pathobiology of the disease.

The second most significant focal CNA was a chromosome 11p13 regional gain targeting genes CD44 and PDHX, present in 17 of 46 cases (37%; Fig.2A–C). CD44 is a nonkinase transmembrane glycoprotein that is induced in B cells upon antigen-mediated activation and is critically involved in multiple lymphocyte functions, including migration, homing, and the transmission of signals that regulate apoptosis. This CD44 protein is also thought to increase cancer cells' adaptive plasticity in response to the microenvironment, thus giving them a survival and growth advantage [310]. Analysis of 20 samples with available RNA-seq data revealed that CD44 was consistently and highly expressed in cases harboring copy number gains (n = 2), whereas PDHX levels were undetectable, indicating that CD44 is the specific target of the 11p13 amplicon. However, all samples displayed high CD44 mRNA expression, independent of the presence of genetic aberrations (Fig.2C); consistently, IHC analysis with a specific CD44 antibody showed very strong membranous staining in all cases tested (Fig.2D) and confirmed this finding

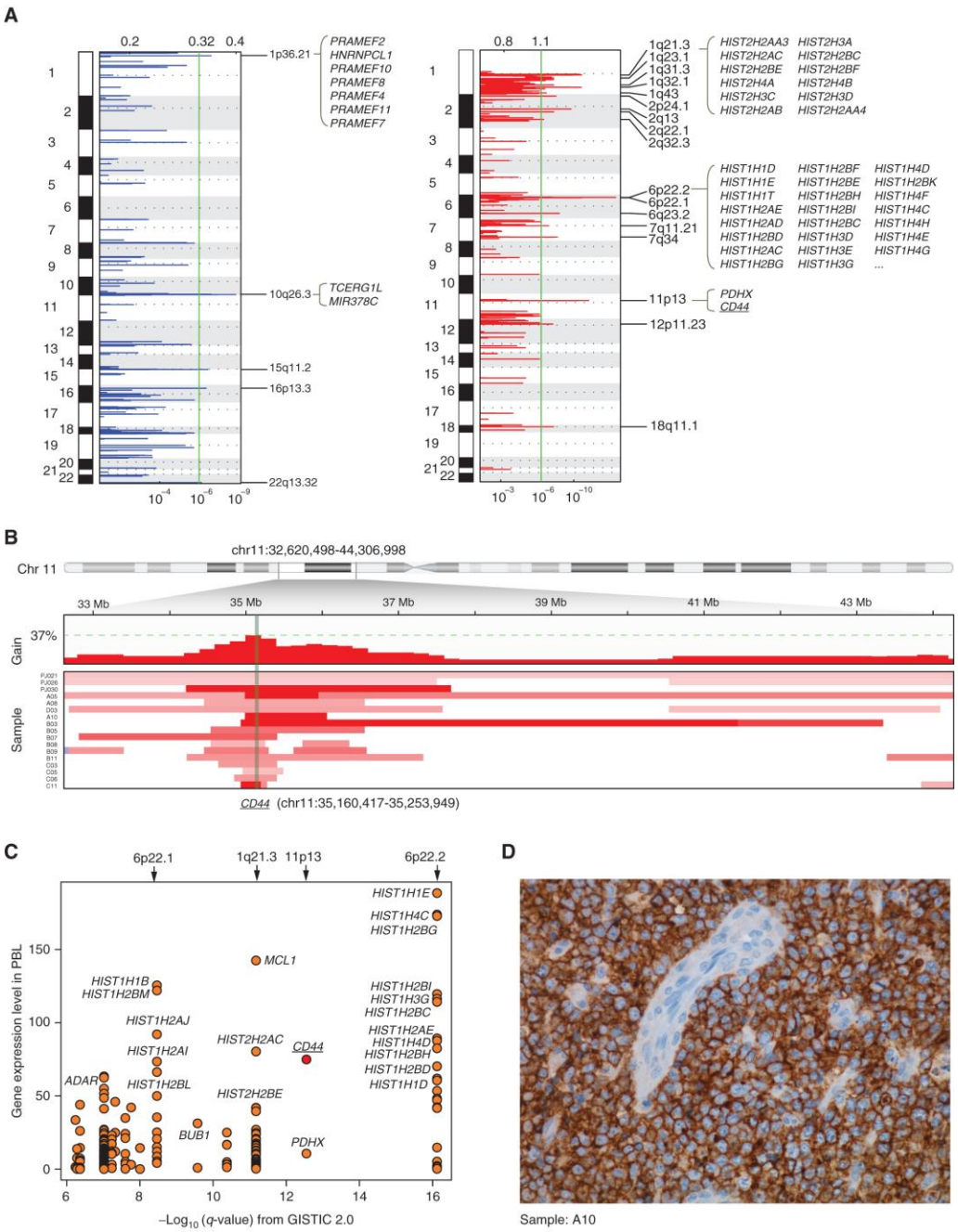


Figure 2: Recurrent copy number changes in PBL. A, GISTIC 2.0 results showing recurrent copy number changes in PBL samples. The green line indicates q-value = 1.0 × 10−6. B, A zoomed-in view of 11p13 on 17 cases of PBL, which shows consistent focal copy number gains of CD44. The figure was generated by the IGV browser using CNV segment files from SNP-FASST2 algorithm. C, Scatter plot representations of genes located in regions with recurrent copy number gains in PBL (q-value <1.0 × 10−6, GISTIC 2.0). The horizontal axis indicates the −log10(q-value) from the GISTIC report, and the vertical axis is the median gene expression level (normalized RPKM value) from PBL RNA-seq data (n = 20). D, IHC showing strong CD44 protein expression on tumor cells membrane in a representative sample (A10), which has a focal CD44 copy number gain.

in an independent panel of 38 cases. Of note, although CD44 expression can be detected in normal plasma cells at both RNA and protein levels, the signal was markedly lower than in plasmablastic lymphoma cells, suggesting that elevated expression of CD44 does not simply reflect the cellular ontogeny of these tumors, and that alternative regulatory mechanisms may lead to CD44 upregulation in cases lacking CNAs (Supplementary Fig. S5A–S5C; Supplementary Table S8).

Plasmablastic Lymphoma Displays a Distinct Genetic and Transcriptional Program

To explore the genomic features of PBL in relation to other lymphoid neoplasms arising from the mature B-cell lineage, we performed unsupervised clustering analysis based on mutation frequency of the top mutated genes from three cancer types obtained from public repositories (Fig. 3A). The analysis included chronic lymphocytic leukemia (CLL; ref. [311]), diffuse large B-cell lymphoma (DLBCL) transcriptionally defined as ABC and germinal center B-cell-like (GCB) subtypes [79], and multiple myeloma [312]. As expected on the basis of their presumed derivation from B cells committed to plasma cell differentiation, the mutational landscape of PBL was overall closer to multiple myeloma than to other mature B-cell malignancies, with mutations in RAS family members being detected in as many as 20% of cases in both diseases, while rare in DLBCL (Fig.3A). Conversely, the mutational landscape of PBL was highly distinct from that of both GCB- and ABC-DLBCL (Fig.3A). In particular, mutations affecting the methyltransferase KMT2D and acetyltransferase CREBBP, two among the most commonly mutated genes in DLBCL [313], were absent in plasmablastic lymphoma. ABC-DLBCL–specific mutations such as CD79A/B and MYD88 were also lacking in PBL. Of note, STAT3 was the top mutated gene in our cohort at significantly different frequencies compared with other B-cell malignancies, making it a hallmark of PBL (Fig.3A). The differential genetic landscape of PBL is consistent with its status as a distinct entity among mature B-cell neoplasms.

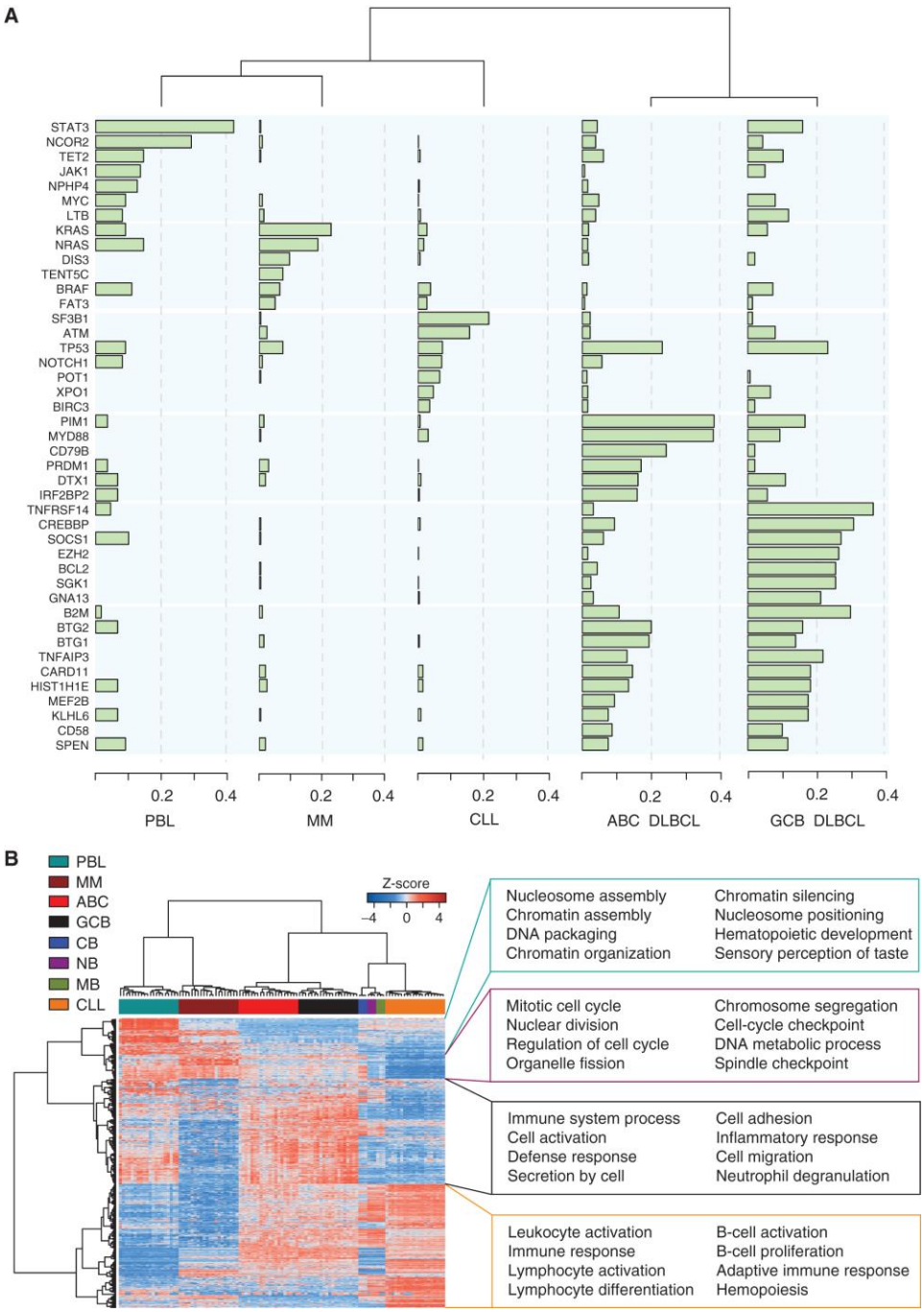


Figure 3: Comparative analysis of PBL and other B-cell malignancies. A, Unsupervised clustering based on frequencies of the most recurrently mutated genes from PBL, multiple myeloma, CLL, and two main subtypes of DLBCL (Methods). B, Hierarchical clustering of mRNA expression profiles across plasmablastic lymphoma, multiple myeloma, CLL, DLBCL and normal B cells, including centroblast (CB), naïve B (NB), and memory B cells (MB). Note that the JAK–STAT signaling pathway does not appear in this figure because only the top 1,000 most aberrantly expressed genes were selected for this analysis. Functional enrichment analysis was performed using g:profiler.

To define the transcriptional profile of HIV-positive PBL and to identify unique signatures that may distinguish it from other lymphoma types, we performed RNA-seq analysis of 20 PBL samples (including 12 of the 15 discovery cases), and compared their transcriptome to that of normal B-cell subsets and other B-cell lymphoma types previously characterized in our laboratories and/or obtained from public repositories, including germinal center centroblasts (CB), naïve B cells (NB), and memory B cells (MB; ref. [314]), as well as CLL [315], DLBCL [79], and multiple myeloma cell lines [316]. As expected, hierarchical clustering of the top 1,000 most aberrantly expressed genes revealed that PBL and multiple myeloma were closer to each other compared with other B-cell malignancies and to normal B cells, reflecting their presumed cell of origin (Fig.3B). Consistently, plasmablastic lymphoma and multiple myeloma lacked expression of common B-cell markers (CD19, CD20, CD40, and PAX5) and transcription factors involved in the germinal center reaction (BCL6, BCL7A, BCL11A, and SPIB), whereas expression of the master regulator of plasma cell differentiation PRDM1 and other plasma cell markers (CD138, XBP1, and IRF4) were increased (Supplementary Fig. S6). MYC expression was also higher in PBL and multiple myeloma. Other notable differences included the upregulation of IL6R, a known STAT3 target, and the downregulation of SWAP70, previously suggested as a potential biomarker of PBL (Supplementary Fig. S6; ref. [298]). Finally, recent studies have suggested that EBV positive PBLs evade immune recognition by expressing the programmed cell death protein 1 (PD-1) and its PD-L1 ligand [317]. We confirmed high expression of PD-L1 in our PBL cohort, although PD-1 did not follow this pattern (Supplementary Fig. S6).

We then performed functional enrichment analysis (g:profiler) to identify biological pathways that are preferentially enriched in PBL as compared with other B-cell malignancies. Whereas DLBCL and CLL were enriched for B-cell related biological programs (e.g., B-cell

activation and proliferation, lymphocyte activation and differentiation, and adaptive immune response), multiple myeloma was enriched for mitotic cell-cycle processes, with highly expressed genes including MCM10, BIRC5, CENPE, BUB1, and AURKA (Fig.3B). The most significantly enriched pathways in PBL were related to chromatin/nucleosome assembly (Fig.3B); particularly, histone genes encoding basic nuclear proteins were highly expressed in PBL, possibly related to the recurrent copy number gains/amplification encompassing this gene cluster on chromosomes 6p22.2 and 1q21.3 (Supplementary Fig. S7A–S7D). However, the biological significance of this observation in relation to tumorigenesis remains unknown.

Activation of the JAK–STAT Pathway in Plasmablastic Lymphoma

Given the elevated frequency of mutations targeting the JAK–STAT pathway, we used computational and IHC approaches to assess the extent of activation of this signaling cascade in PBL (Methods). To this end, we first interrogated the genome-wide transcriptional profile of PBL for enrichment in the JAK–STAT signaling pathway using previously identified signatures available in the MSigDB database. This analysis revealed a significant positive enrichment for genes implicated in this pathway, consistent with constitutive activation (Fig.4A). This signature was expressed at higher levels in PBL when compared with multiple myeloma and CLL (Fig.4B), with the most significant difference being observed between PBL and multiple myeloma (t test, $P = 4.7 \times 10{-}11$), which is congruent with the fact that the latter had few mutations in this pathway (Fig.3A). We then performed IHC staining with anti-STAT3 and anti-phospho-STAT3 antibodies to assess the STAT3 subcellular localization and phosphorylation, used as a read-out for signaling activation. We observed strong positive nuclear signal in 61% of tested cases (19/31), which also stained positive for the phospho-STAT3 antibody. Of note, evidence of constitutively active STAT3 was detected even in the absence of genetic alterations affecting this pathway, suggesting

alternative (epigenetic) mechanisms of activation and indicating a prominent role for this cascade in the pathogenesis of the disease (Fig.4C and D; Supplementary Fig. S8; Supplementary Table S9).

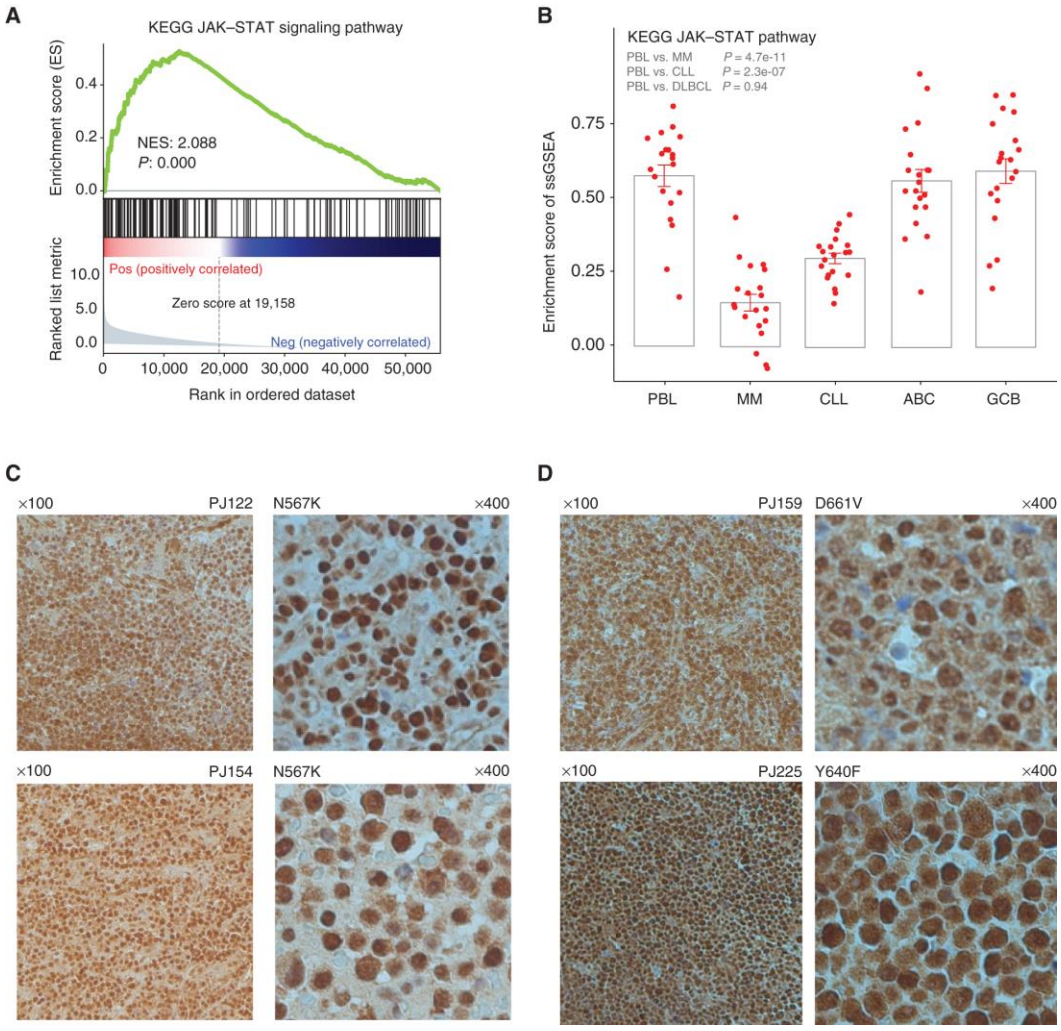


Figure 4: Activation of JAK–STAT pathway in PBL. A, Preranked gene set enrichment analysis indicating the significant positive enrichment of KEGG JAK–STAT signaling pathway in PBL. The preranked gene list was generated on the basis of the median expression level on PBL samples. B, Enrichment comparisons of JAK–STAT pathway between PBL and other B-cell malignancies by performing single-sample GSEA (ssGSEA). Pairwise P values were derived from t test. C and D, IHC plots showing pSTAT3 protein expression in >75% of tumor cells, confirming STAT3 activation in STAT3 mutated cases. Results using anti-pSTAT3 antibody are photographed at ×100 and ×400 magnification.

Virus Detection in Plasmablastic Lymphoma

To analyze the viral and bacterial make-up of PBL tumors, we used the Pandora pipeline, which extracts and aligns nonhost genetic material from tumor RNA-seq data (Methods). Potentially pathogenic species were then identified by applying the BLAST algorithm against the NCBI database of viruses and bacteria reference genomes. In our PBL cohort, 12 of 20 samples contained HIV-1 transcripts, with only three samples exceeding 100 mapped reads (Supplementary Fig. S9). In addition, EBV transcripts were detected in 18 of 20 samples, HCMV (human cytomegalovirus, or human betaherpesvirus 5) in 3 of 20 samples, and KSHV (Kaposi sarcoma-associated herpes virus, human herpes virus 8) in one sample (Supplementary Fig. S9). These three herpes viruses have a broad tropism, naturally infect B cells, and are known to be associated with many tumor types.

EBV reactivation is considered a major driver of PBL, predominantly with a latency I infection program although low levels of LMP1 gene expression can be detected [318]. We thus performed a detailed RNA profiling of the EBV genome. Recapitulating previous reports, all PBL cases tested by in situ hybridization were positive for EBV-encoded RNA (EBER), while only 4 of 13 samples showed expression of the LMP1 protein on IHC (Supplementary Table S1). When using the levels of BLLF1 and EBNA2, two genes that are invariably not expressed in PBL, as threshold for positive calls, we noted that nearly the entire viral genome was transcribed at background levels in the majority of samples (Fig.5). This is analogous to what has been observed in nonreplicating infected B cells, where a background of transcripts can be detected in the absence of protein expression, due to regulation at the ribosomal level [319]. However, several genes in the BamHI-A region of the virus were abundantly transcribed in at least 50% of cases across the cohort, including those encoding for components of the viral replication machinery (namely, the

DNA polymerase catalytic subunit BALF5, the single-stranded DNA-binding protein BALF2, and the lytic origin of DNA replication oriLyt), the BART encoded protein RPMS1, the viral envelope glycoprotein BALF4, and the G-protein–coupled receptor BILF1, which is expressed predominantly during the immediate early phases of infection in vitro [320, 321]. LF1, LF2, and LF3 were also highly expressed in PBL, but their functional roles are less understood. Expression of the LMP-1 mRNA was detected in 17 of 20 samples, but at much lower levels (Fig.5), consistent with the known EBV latency programs of PBL [318].

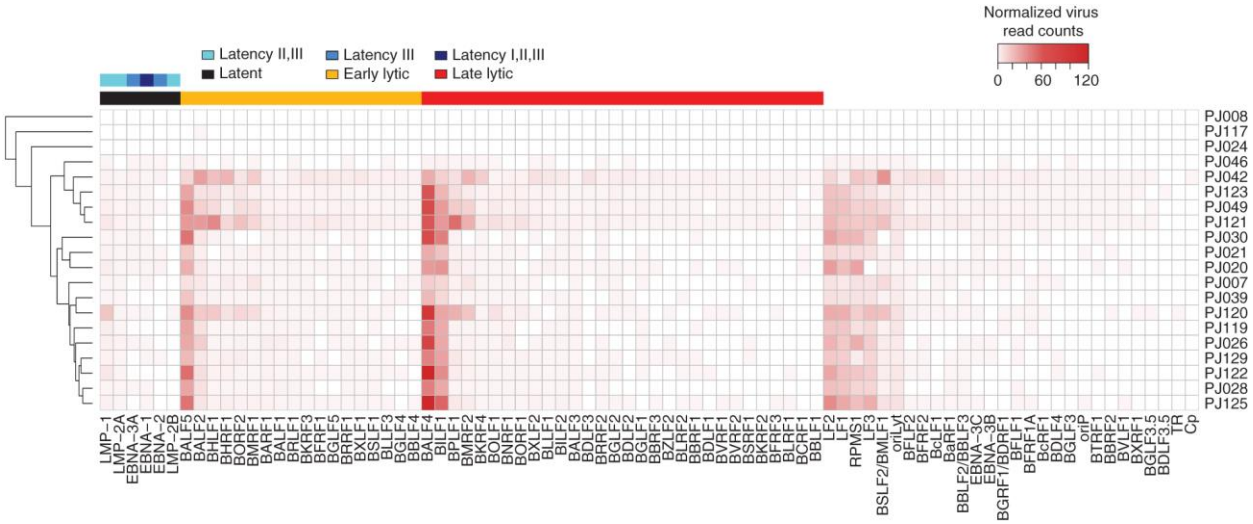


Figure 5: EBV transcription programs in PBL. Heatmap illustrating the full genome expression of EBV in the RNA-seq data of PBL samples (virus gene counts are normalized per million host reads). 16/17 EBV-positive samples show significantly higher expression of lytic genes such as *BALF4* (encoding the envelope glycoprotein B) and *BALF5* (encoding the DNA polymerase catalytic subunit) compared with the expression of canonical latency programs.

A.3 Discussion

Our study provides a comprehensive snapshot of the genetic landscape of HIV-associated PBL and reveals highly recurrent somatic mutations affecting the JAK–STAT3 and RAS–MAPK signaling pathway as a genetic hallmark of this disease, with STAT3 representing the most prominent target. These findings underscore a central role for this transcription factor in the

pathogenesis of PBL and have implications for the diagnosis and treatment of these diseases.

The STAT3 protein is an important player in multiple immune cells where it modulates a variety of physiological processes. Within the B-cell lineage, a selective role has been recognized for this factor in the differentiation of B cells into plasma cells upon antigen stimulation, as documented by both in vitro and in vivo studies [300]. Briefly, in response to CD40L- and IL21-mediated signaling by T follicular heper cells, STAT3 is phosphorylated by activated JAK kinases, translocates into the nucleus as homo- or hetero-dimers, and activates the transcription of multiple targets, including the plasma cell master regulator BLIMP1. In turn, BLIMP1 downregulates BCL6 expression, an absolute prerequisite to GC exit and plasma cell differentiation [300]. The STAT3 amino acid changes identified in our study include experimentally documented gain-of-function events that are predicted to have oncogenic effects by enhancing its phosphorylation and transactivation potential [322]. In addition, other well-documented genetic mechanisms were found that can activate the JAK–STAT signaling pathway, including mutually exclusive upstream mutations of JAK1 or JAK2 (18/110 cases) and the loss of the STAT3 negative regulator SOCS1 (mutated in 11/110 cases). Notably, a targeted sequencing study published while this manuscript was under revision identified recurrent STAT3 mutations in 5 of 42 cases, all of which were EBV positive [323]. Thus, STAT3 dysregulation may contribute to the pathogenesis of PBL by rendering these cells signaling-independent while providing proliferation and survival signals.

Constitutive activation of the JAK–STAT signaling pathway has been reported in a number of solid and hematologic malignancies and plays a central role in two lymphomas that are immunophenotypically closely related to plasmablastic lymphoma: primary effusion lymphoma (PEL) and ALK-positive large B-cell lymphoma (LBCL). PEL is a rare and aggressive AIDS-defining disease, which is associated with infection by HHV8 and is clinically distinguishable from

PBL by the presence of lymphomatous effusion in body cavities. In these cells, constitutive STAT3 activity is achieved via expression of the HHV8 viral protein IL6, which contributes to the disease in an autocrine fashion by promoting proliferation and survival [324]. In ALK-positive LBCL, STAT3 activation is sustained by the ALK kinase mediated by chromosomal translocations with the CLTCL gene or the NPM1 gene [325, 326]. However, direct genetic alterations of STAT3 are rare in mature B-cell lymphomas. In particular, while multiple genomic hits leading to potentiation of the JAK–STAT oncogenic pathway have been detected in 87% of Hodgkin lymphomas as well as in primary mediastinal B-cell lymphomas, the most commonly affected STAT member in these tumors is STAT6 [93, 327-329]. The high incidence of STAT3 mutational activation in HIV-associated PBL points to STAT or JAK inhibitors as promising treatment options in this lymphoma type. While anti-STAT3 therapeutic attempts are still in development, JAK inhibitor therapy, currently used in the clinical setting, was shown to be an effective antagonist to STAT3 activation, inducing apoptosis in both anaplastic large T-cell lymphomas and ovarian cancer [330].

The second important finding of this study is the identification of frequent hotspot mutations in RAS–MAPK family members. Functional RAS activation is a common molecular feature of multiple myeloma, particularly in the relapsed/refractory setting, while it is rarely observed in *de novo* DLBCL, suggesting a specific role in the pathogenesis of plasma cell dyscrasias. Interestingly, and different from multiple myeloma, NRAS, and KRAS mutations in PBL were never concurrently observed in the same case and were often well represented in the dominant tumor clone, consistent with early events. These data have direct implications for the clinical exploration of treatments inhibiting this pathway.

Our study also showed overexpression of the transmembrane glycoprotein CD44, frequently associated with copy number gains/amplifications at this locus. CD44 is an adhesion

molecule that mediates cellular interaction with the microenvironment and participates in the trafficking of neoplastic cells in multiple myeloma, CLL, and ALL [331]; moreover, CD44 was shown to increase cell resistance to apoptosis and to enhance cancer cell invasiveness. Whereas the role of CD44 in PBL requires functional dissection, its high expression is likely to compound the aggressiveness of the disease, as previously described in DLBCL of the ABC subtype [332]. In light of this, successful anti-CD44–targeted therapy in a mice xenograft model of human multiple myeloma may in future represent an attractive therapeutic option for PBL [333].

The finding of increased histones mRNA abundance in PBL, together with recurrent copy number gains encompassing this gene cluster on chromosome 6, is of interest because it emerged as a distinctive feature of this disease compared with other lymphoid malignancies. Histones represent a basic component of the chromatin structure and their involvement may suggest a selective role for nucleosomal plasticity in the pathogenesis of PBL, which warrants further investigations.

Due to the small number of EBV-negative cases in our cohort (n = 2) and the overall rarity of this subset, we could not assess whether EBV status is associated with differing genetic features, as recently reported for Burkitt lymphoma [249] and suggested by the evidence of distinct transcriptomic profiles between EBV-positive and –negative PBL [317]. Moreover, it remains to be determined whether regional differences exist with PBL occurring outside the context of HIV immunodeficiency, or within HIV-associated PBL, given the high homogeneity of our cohort from the Gauteng region in South Africa. Thus, additional work will be required to specifically address these questions.

We found that most of the EBV genome was transcribed at very low levels in the majority of PBL cases, while prominent expression was detected for several genes in the BamHI-A region

of the virus, including some early lytic genes. However, transcription of the key early lytic gene BZLF1 was noticeably absent, suggesting an incomplete lytic program. Supporting this notion, a recent study showed that increased STAT3 expression, as observed in the majority of PBL cases, decreases the susceptibility of latently infected cells to EBV lytic activation signals via an RNA-binding protein PCBP2 [334]. Thus, further protein expression studies, in particular for BRLF1 and for the BLLF1-encoded viral envelope glycoprotein gp350, are required to assess whether the observed transcriptome program translates in lytic infection. Indeed, recent translation ribosome profiling studies have clearly demonstrated a marked heterogeneity of lytic genes translation and complex levels of intracellular translation repression mechanisms at work in infected B cells [319].

In conclusion, the results of our study characterize HIV-associated PBL as a distinct subset of aggressive B-cell lymphoma and significantly contribute to our knowledge about the molecular pathogenesis of this disease through the identification of recurrently mutated genes, uncovering a major role for the dysregulation of JAK–STAT3 and RAS–MAPK signaling pathways. These results reveal new points of potential therapeutic intervention in these patients.

A.4 Methods

For complete experimental details and computational analyses, see also Supplementary Methods.

Patient Cohorts

Prospective samples (n = 15 with paired tumor-normal tissue; discovery cohort) were collected from patients with suspected PBL, upon informed written consent in line with the Declaration of Helsinki, and diagnosis was further confirmed by two independent pathologists. Matched normal DNA for these 15 patients was extracted from peripheral blood samples that were documented to be tumor-free. For targeted sequencing (n = 95 samples; extension cohort), formalin-fixed paraffin-embedded (FFPE) material was retrieved from the archives of the

Department of Oral Pathology and Anatomical Pathology, National Health Laboratory Service and the University of the Witwatersrand. This panel included 36 nonoral PBL samples that were part of a published cohort [335]. The study protocol was approved by the local Human Research Ethics Committee (IRB Reference M150390). A summary of demographics and phenotypic markers of the discovery cohort are displayed in Supplementary Table S1, whereas demographic information for the extension cohort is included in Supplementary Table S3. For downstream nucleic acid extraction, the tumor area (>70% tumor cells) was ringed for microdissection in all samples. A third cohort of 31 well-characterized PBL samples obtained as archived FFPE material (local Human Research Ethics Committee IRB reference M101171 and 96/2011; Supplementary Table S7) was used to validate and refine copy number aberration results observed in the WES data by using Microarray OncoScan [307].

WES and RNA-Seq

Both exome and RNA-seq library preparations and sequencing were outsourced to Centrillion Genomics Services and BGI (AmericasCorporation). Whole-exome libraries were prepared using the Agilent SureSelect Human All Exon V6 Kit (Agilent Technologies) and sequenced on HiSeq 2500 using TruSeq SBS v2 Reagent Kit (Illumina) at 2 × 100 bp paired-end reads with on-target coverage of 100X per sample. RNA-seq libraries were prepared using the Illumina TruSeq Stranded Total RNA Sample Prep Kit (Illumina). Flow cells with multiplexed samples were run on the HiSeq 2500 using an Illumina TruSeq SBS v2 Reagent Kit at 2 × 100 bp paired-end reads and a coverage of 50M reads per sample.

Mutation Calling

Fastq files were aligned to the human genome assembly (hg19) using the Burrows–Wheeler Aligner (version 0.6.2). Before further analysis, the initially aligned BAM files were

subjected to preprocessing that sorted, indexed, and marked duplicated reads using SAMtools (version 1.7) and Picard (version 1). To identify somatic mutations from WES data for tumor samples with matched blood control, we applied the variant-calling software SAVI2, based on an empirical Bayesian method as published [122]. Somatic mutations were identified on the basis of the final report of SAVI2, and following five additional criteria: (i) not annotated as a synonymous variant, intragenic variant, or intron variant; (ii) not annotated as a common SNP (dbSnp138); (iii) a variant allele frequency of >5% in the tumor sample; (iv) a variant allele read depth of <2 in the matched normal control; and (v) variant associated reads with overall mismatch rate ≤ 0.02 as estimated by bam-readcount (version 0.8.0, https://github.com/genome/bam-readcount). All mutations described throughout this manuscript refer to nonsynonymous mutations, unless otherwise specified.

Design of Targeted Sequencing Panel

The full coding exons of 34 genes found recurrently mutated in the discovery cohort and/or previously implicated in lymphoma were analyzed in an extension panel of 95 cases (Supplementary Table S3) by targeted capture and next generation sequencing. Genes were selected according to the following criteria: (i) allele frequency > 15%; (ii) mutated in at least 2 of 15 cases; (iii) expressed in normal or transformed B cells; and (iv) functionally annotated. In addition, we included 16 genes that were only mutated in one sample but have known roles in the pathogenesis of lymphoma and 4 genes that were not found mutated in the discovery cohort but have been previously implicated in PBL. Genes lacking clear functional annotation and/or known to represent common nonspecific mutational targets in sequencing studies (e.g., TTN, PCLO) were excluded. The complete gene list is reported in Supplementary Table S4.

Targeted Next-Generation Sequencing

The entire coding region of the 34 selected genes was isolated using the IDT xGen Predesigned Gene Capture Custom Target Enrichment Technology (Integrated DNA Technologies) and subjected to library preparation and next-generation sequencing on the Illumina HiSeq platform with 2 × 150 bp configuration. Targeted sequencing was performed at GENEWIZ. Read alignments and conventional preprocessing were conducted as described for the WES analysis. For samples lacking matched normal control, variants were filtered out if found in dbSNP database (dbSNP138) as well as in any normal sample of the WES cohort.

Copy Number Analysis

For copy number analysis from WES data (n = 15), the Biodiscovery Multiscale BAM Reference Builder [336] was used to construct a multiscale reference (MSR) file from 14 paired normal samples alignments (BAM). The MSR file was used as reference for copy-number variation (CNV) calling of all tumor alignments with the SNP-FASST2 algorithm [337], using Nexus Copy Number, v10.0 (BioDiscovery, Inc.; ref. [337]). Gains and losses were defined as at least +0.3 and -0.3 log2 ratio changes, respectively, in the tumor alignment.

To validate CNV calling from WES data, 31 additional samples were processed on Oncoscan FFPE Express Arrays (Affymetrix, Thermo Fisher Scientific) (Supplementary Table S7) according to the manufacturer's instructions, followed by scanning on a GeneChip Scanner 3000 7G, with the Affymetrix GeneChip Command Console (AGCC). Paired A+T and G+C CEL files were combined and analyzed with Chromosome Analysis Suite v3.3.0.139 (ChAS, Applied Biosystems, Thermo Fisher Scientific), using the OncoScan CNV workflow for FFPE, without manual recentering. We confirmed that copy number calls from OncoScan and WES data were consistent with each other by performing OncoScan on 7 cases that had been sequenced by WES

and comparing the calls obtained from the two methods. Probe genomic coordinates were aligned to hg19 and the resulting OSCHP files were analyzed by Nexus Copy Number v10.0 using the SNP-FASST2 algorithm [337] with default parameters.

To detect recurrent copy number aberrations, we applied the GISTIC 2.0 algorithm using GenePattern (https://www.genepattern.org/) to the copy number segmentations of the combined cohort of 46 patients (15 cases analyzed by WES, 31 cases analyzed by OncoScan). Recurrent regions of copy number aberrations with q-value $< 1.0 \times 10{-}6$ were considered significant. Furthermore, we assessed the expression level of each gene within the significant GISTIC peaks using the median value of quantile normalized RPKM from RNA-seq data (n = 20).

Pandora: A High-Performing Pipeline for Quantifying the Bacterial and Viral Microenvironment of Bulk RNA-Seq Samples

Pandora is an open-source pipeline (https://github.com/RabadanLab/Pandora) that takes as input the total RNA sequence reads from a single sample and outputs the spectrum of detected microbial transcripts, focusing on bacteria and viruses. The Pandora workflow is divided into the following four main modules: (i) mapping to the human host genome using STAR and Bowtie2 to filter out the host reads from downstream analysis; (ii) *de novo* assembly of host-subtracted short reads using Trinity [338] to create contiguous full-length transcripts (contigs), which help increase the accuracy of alignment to the correct species of origin in the next step; (iii) identification of the most likely species of origin for each assembled contig with BLAST; and (iv) filtering and parsing of the BLAST results into a final report on the detected microbial abundances. We also performed gene expression profiling of all the reads that mapped to the EBV genome [339] using FeatureCounts [340] to fully characterize lytic and latent EBV programs.

Data Availability

The data that support the findings of this study are available upon request. The sequencing data have been deposited in NCBI Sequence Read Archive (SRA) under accession number PRJNA598849.

A.5 Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

A.6 Authors' Contributions

Z. Liu: Conceptualization, investigation, visualization, methodology, writing-original draft, writing-review, and editing. I. Filip: Investigation, methodology, and writing-original draft. K. Gomez: Investigation, methodology, writing-review, and editing. D. Engelbrecht: Resources, validation, investigation, and visualization. S. Meer: Resources and investigation. P. Lalloo: Resources, validation, and investigation. P. Patel: Resources, validation and investigation.Y. Perner: Resources and investigation. J. Zhao: Investigation, visualization, and methodology. J. Wang: Investigation, writing-review, and editing. L. Pasqualucci: Conceptualization, resources, supervision, funding acquisition, investigation, methodology, writing-original draft, project administration, writing-review, and editing. R. Rabadan: Conceptualization, supervision, funding acquisition, investigation, methodology, project administration, writing-review, and editing. P. Willem: conceptualization, resources, supervision, funding acquisition, validation, investigation, visualization, writing-original draft, project administration, writing-review, and editing.

A.7 Acknowledgements

This work has been funded by NIH grants R21 CA192854 (to P. Willem, L. Pasqualucci, and R. Rabadan), R01GM117591 and U54 CA193313 (to R. Rabadan), and was initiated under the Columbia-South Africa Training Program for Research on HIV-associated Malignancies D43

CA153715, with the support of Judith Jacobson. We thank Stephen P. Goff and Henri-Jacques Delecluse for their helpful suggestions. We also thank Sonja Boy for providing additional samples for the independent copy number validation cohort, Nicole Crawford for assistance in samples collection and data curation, and Jacky Brown for help with Sanger sequencing. Whole exome capture and sequencing, and RNA sequencing were completed at Centrillion Biosciences, Inc and BGI Tech Solutions (Hong Kong). Targeted DNA sequencing was performed at Genewiz Inc. **Note:** Supplementary data for this article are available at Blood Cancer Discovery Online (http://bloodcancerdiscov.aacrjournals.org/).