DNA methylation epitypes of Burkitt lymphoma with distinct molecular and clinical features

Nicole Thomas[1,†], Carlos A. Garcia-Prieto[2,3,†], Kostiantyn Dreval[1,4], Laura K. Hilton[1,5], Jeremy S. Abramson[6], Nancy L. Bartlett[7], Jeffrey Bethony[8], Jay Bowen[9], Anthony C. Bryan[9], Corey Casper[10], Maureen A. Dyer[11], Julie M. Gastier-Foster[9,12,13], Alina S. Gerrie[5,14], Timothy C. Greiner[15], Nicholas B. Griner[16], Thomas G. Gross[17], Nancy Harris[18], John D. Irvin[19], Elaine S. Jaffe[20], Fabio E. Leal[21], Sam M. Mbulaiteye[22], Charles G. Mullighan[23], Andrew J. Mungall[4], Karen L. Mungall[4], Constance Namirembe[24], Ariela Noy[25], Martin D. Ogwang[26], Jackson Orem[24], German Ott[27], Hilary Petrello[9], Steven J. Reynolds[28], Steven H. Swerdlow[29], Alexandra Traverse-Glehen[30], Wyndham H. Wilson[31], Marco A. Marra[4,5,32], Louis M. Staudt[31], David W. Scott[5,14], Manel Esteller[2,33,34,35*], Ryan D. Morin[1,4,5*]

[1] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada
[2] Cancer Epigenetics Group, Josep Carreras Leukaemia Research Institute (IJC), Badalona, Barcelona, Catalonia, Spain
[3] Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Catalonia, Spain
[4] Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada
[5] Centre for Lymphoid Cancer, BC Cancer, Vancouver, BC, Canada
[6] Center for Lymphoma, Massachusetts General Hospital, Boston, MA, USA
[7] Washington University School of Medicine, St. Louis, MO, USA
[8] George Washington University, Washington, DC, USA
[9] Biopathology Center, Nationwide Children's Hospital, Columbus, OH, USA
[10] Access to Advanced Health Institute, Seattle, WA, USA
[11] Clinical Research Directorate, Frederick National Laboratory for Cancer Research, Frederick, MD, USA
[12] Department of Pathology and Laboratory Medicine, Ohio State University, Columbus, OH, USA
[13] Departments of Pediatrics and Pathology & Immunology, Baylor College of Medicine, Houston, TX, USA
[14] Division of Medical Oncology, Department of Medicine, University of British Columbia, Vancouver, BC, Canada
[15] Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE, USA
[16] Office of Cancer Genomics, National Cancer Institute, Bethesda, MD, USA
[17] Center for Global Health, National Cancer Institute, NIH, Rockville, MD, USA

[18] Department of Pathology, Massachusetts General Hospital, Boston, MA, USA
[19] Foundation for Burkitt Lymphoma Research, Geneva, Switzerland
[20] Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[21] Programa de Oncovirologia, Instituto Nacional de Cancer, Rio de Janeiro, Brazil
[22] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Silver Spring, MD, USA
[23] Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN, USA
[24] Uganda Cancer Institute, Kampala, Uganda
[25] Department of Medicine, Lymphoma Service, Memorial Sloan Kettering Cancer Center, and Weill Cornell Medical School, New York, NY, USA
[26] St. Mary's Hospital, Gulu, Uganda
[27] Department of Clinical Pathology, Robert-Bosch-Krankenhaus, and Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, Germany
[28] Division of Intramural Research, National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA
[29] Division of Hematopathology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
[30] Service d'Anatomie pathologique, Université Lyon 1/Centre Hospitalier Lyon Sud, Pierre Bénite, France
[31] Lymphoid Malignancies Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[32] Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada
[33] Centro de Investigacion Biomedica en Red Cancer (CIBERONC), Madrid, Spain
[34] Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
[35] Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Barcelona, Catalonia, Spain

[†]These authors contributed equally

Running title

Burkitt lymphoma epitypes

***Co-corresponding Authors**

Dr. Ryan D. Morin

Department of Molecular Biology and Biochemistry

Simon Fraser University

8888 University Dr.
Burnaby, BC, Canada
V5A 1S6
rdmorin@sfu.ca

Phone: +1 778 782 5630

Dr. Manel Esteller

Josep Carreras Leukaemia Research Institute

Carretera de Can Ruti, Camí de les Escoles s/n

08916 Badalona, Barcelona, Catalonia, Spain

mesteller@carrerasresearch.org

Phone: +34 935 572 837

Conflict of Interest Statement

R.D.M. and D.W.S. are named inventors on a patent application describing the double-hit signature. C.G.M. received research funding from Pfizer, AbbVie; advisory board member at Illumina and speakers bureau at Amgen. A.N. received research funding from Pharmacyclics/AbbVie, Kite/Gilead, and Cornerstone; was a consultant for Janssen, Morphosys, Cornerstone, Epizyme, EUSA Pharma, TG Therapeutics, ADC Therapeutics, and Astra Zeneca; and has received honoraria from Pharmacyclics/AbbVie. M.E. reports grants from Incyte and personal fees from Quimatryx outside the submitted work. All other authors declare no competing financial and/or non-financial interests in relation to the work described.

Abstract

The genetic subtypes of Burkitt Lymphoma (BL) have been defined, whereas the role of epigenetics remains to be comprehensively characterized. We searched genomic DNA from 218 patients across four continents, for recurrent DNA methylation patterns and their associations with clinical and molecular features. We identified DNA methylation patterns that were not fully explained by EBV status or mutation status, leading to two epitypes, described here as HypoBL and HyperBL. Each is characterized by distinct genomic and clinical features including global methylation, mutation burden, aberrant somatic hypermutation, and survival outcomes. Methylation, gene expression and mutational differences between the epitypes support a model in which each arises from a distinct cell-of-origin. These results, pending validation in external cohorts, point to a refined risk assessment for BL patients who may experience inferior outcomes.

Statement of significance

Burkitt lymphoma (BL) can be divided into two epigenetic subtypes (epitypes), each carrying distinct biological, transcriptomic, genomic, and clinical features. Epitype is more strongly associated with clinical and mutational features than EBV status or genetic subtype, highlighting an important additional layer of BL pathogenesis.

Introduction

Burkitt Lymphoma (BL) is the most prevalent type of B-cell non-Hodgkin lymphoma (NHL) affecting children, accounting for approximately 50% of pediatric NHL cases whereas it is more rare in adults(1–3). The genetic hallmark is a chromosomal translocation that places *MYC* under the regulation of a potent enhancer, causing constitutive expression(4,5). BL has historically been classified based on clinical variant status (endemic or sporadic) and patient age (pediatric or adult)(6,7). Recent evidence has underscored the association of Epstein-Barr virus (EBV) tumor positivity with distinct genomic features(8–10). These insights have challenged the relevance of clinical variant status, pointing to EBV status as more biologically meaningful. Using genome-wide mutational information, four genetic subgroups of BL have been proposed, which span pediatric and adult BL. These are marked by distinct combinations of driver mutations, namely in *DDX3X*, *GNA13*, *GNAI2* (DGG-BL), *ID3*, *CCND3* (IC-BL), and *TP53* (Q53-BL), and some of them bearing distinct patterns of non-coding mutations caused by aberrant somatic hypermutation (aSHM)(9).

The genome-wide methylation features of BL have yet to be extensively studied. Changes to CpG methylation in tumors typically involves widespread/global hypomethylation and local hypermethylation of specific genes. Specifically, changes to promoters and other regulatory regions can influence expression of tumor suppressor genes(11,12). This tends to preserve some of the methylation pattern of the founding cell and this epigenetic memory can inform on the cell-of-origin(13–17). Owing to different natural histories, differentiation stages, and selective pressures experienced by these tumors, DNA methylation patterns may also be associated with different driver mutations(16–18). Considering the relatively low number of BL driver genes relative to other mature B-cell neoplasms, we sought to explore the contribution of epigenetic changes to BL and how these might inform on other clinical or molecular features.

Results

Identifying and distinguishing the epigenetic landscape of BL

We applied array-based genome-wide methylation profiling to 218 BL biopsies collected at diagnosis (156 pediatric and 62 adult) along with 6 normal centroblast samples**.** We also performed whole genome sequencing on nine of these samples using the PromethION platform and used these data to resolve CpG methylation sites and entropy . Our analysis leveraged existing genomic and clinical information from the same cohort, which was described previously(8,9). As has been observed in other cancers, the BLs exhibited widespread DNA hypomethylation with localized hypermethylation relative to centroblasts (11,12) (**Figure 1A-B**). Specifically, this comparison identified 357,124 differentially methylated probes (DMPs) with 191,724 showing hypomethylation and 165,400 showing hypermethylation in BL. When this comparison was performed separately according to EBV status, EBV-positive BLs had more regions affected by methylation changes relative to centroblasts and a significantly higher number of DMPs exhibiting hypermethylation (**Figure 1C-D;** $P<1x10^{-16}$, Fisher's Exact). The number of DMPs in each chromatin state should be approached with caution as the extent of probe coverage in each of these regions is constrained by the array design. It is clear, however, that the majority of BL-associated hypomethylation affects regions that are typically heterochromatin, regardless of EBV status. In contrast, the regions affected by hypermethylation were more diverse in both EBV-positive and EBV-negative BL (**Figure 1D**).

Through region-level analyses, we identified 10,715 and 7,320 differentially methylated regions (DMRs) in EBV-positive and EBV-negative BL samples, respectively (**Supplemental Tables 1-2**). Of these, 49,855 DMPs and 5334 DMRs were shared between EBV-positive and EBV-negative comparisons, suggesting that changes to methylation within these regions is a shared feature across BL (**Figure 1E**). These pan-BL DMRs span 3340 genes, including genes differentially expressed between centrocytes and centroblasts (*CXCR5, CCND1, SLAMF1*), genes hypermethylated in other B-cell lymphomas (*TP73, IL12BR2, GSTP1, MT1G, CDH1, RARB, RBP1, SLIT2,*

DLC1, p16, *DAPK1, KLF4, DBC1*), and genes with differential expression between BL and diffuse large B-cell lymphoma (DLBCL) (*SOX11, STAT3, CD44, CTSH, DLEU1,BATF,BCL2*)(12,19–28)(**Supplemental Table 3**).

We next investigated whether genes commonly mutated in BL(8,9,29) ("BL genes") are also associated with DNA methylation patterns consistent with transcriptional activity. We compared the methylation state of the promoter with the expression of the corresponding gene across all samples. Among these genes, there was a negative correlation between methylation and expression for *CREBBP, DTX1, HIST1H1D, ID3, KLHL6, SMARCA4, TCL1A*, and *TET2* (**Supplemental Figure 1A**). This is consistent with the notion that the expression of these genes is repressed by DNA methylation in a subset of BLs. As the function of a gene could be similarly suppressed genetically, we incorporated the mutation status of each gene into our analysis using a linear model. This confirmed an association between *ID3* promoter methylation and mutation status. Samples carrying *ID3* mutations exhibited lower promoter methylation (mean beta = 0.158) whereas BL with intact *ID3* had a significantly higher degree of methylation (mean beta = 0.199; P<0.001). Although the promoter exhibited a predominantly hypomethylated state across all BL samples (mean beta = 0.182), we noted that subtle differences in *ID3* promoter methylation were associated with significant changes in gene expression (*P*<0.001; **Supplemental Figure 1B**). A similar trend was observed for *SMARCA4*, although for this gene the difference in the expression level did not reach significance (*P*=0.09; **Supplemental Figure 1B**). Taken together, these data support the hypothesis that the loss of *ID3* function can be acquired through promoter hypermethylation as well as genetic loss.

Because non-coding mutations also have the potential to influence expression through the alteration of regulatory sequences, we investigated the methylation of non-coding regions enriched for mutations in BL, specifically those affected by aSHM(9). Given the disparate rates of aSHM observed between EBV-positive and EBV-negative BL, we explored whether methylation patterns associated with EBV status were correlated with frequencies of aSHM(8,9). To compare with DLBCL, which has higher rates of aSHM relative to BL(9), this analysis included array data from 56 DLBCLs from

a previous study(18) . We noted a general trend of low average beta values across most of analyzed aSHM regions, which is seen across BL, DLBCL and 15 normal germinal center (GC) B cells, whereas a minority of these regions exhibited variable levels of methylation among the BLs and DLBCLs (**Supplemental Figure 1C**). The highly methylated regions were consistently observed across the BLs and GC B cells, implying this epigenetic pattern reflects the state of B cells at this differentiation stage. The hypermethylation of *SEPT9* was predominantly observed in EBV-positive BL, consistent with the lower rates of aSHM across this region in EBV-positive BL samples compared to EBV-negative BLs and DLBCLs(9) (**Supplemental Figure 1D**). Additionally, *ST6GAL1* had the highest degree of methylation in EBV-positive BL, where similar rates of aSHM were observed between EBV-positive and EBV-negative BL (**Supplemental Figure 1D**). While most of these hypermethylated regions were also methylated in GC cells, *ST6GAL1* methylation was uniquely observed in BL. Overall, there was no consistent association between methylation state and the degree of mutations in these regions.

EBV-associated DNA methylation patterns in BL

To explore the effect of EBV on the BL methylome, we stratified BLs on EBV status and searched for methylation differences. This identified 30,845 significant DMPs (|abs(logFC)|> 0.25 and Q < 0.05), with most (30,726) hypermethylated in EBV-positive BL. DMPs were primarily located in openSeas (15985), CpG shores (8742), CpG islands (3021), and CpG shelves (2965). Using chromatin and transcriptional state annotations, we found a significant (P < 0.001) enrichment of hypermethylated DMPs in annotated enhancer regions (**Figure 2A**). Using a region-level analysis, we identified a total of 6,000 DMRs, again with most (5,991) hypermethylated in EBV-positive BL (**Figure 2B**, **Supplemental Table 4**).

There are gene expression differences between EBV+ and EBV- BL, which may be explained, in part, by DNA methylation changes orchestrated by the virus. We searched for genes near DMRs with correlation between expression and methylation levels. Among the 4782 genes associated with DMRs, 476 had significant correlations

(|abs(R^2)|> 0.4 and Q < 0.01) (**Figure 2C**, **Supplemental Table 5**). Most (430) of these genes exhibited anticorrelation between methylation and expression level. Among these DMR/gene pairs, we identified several genes with relevance in B-cell lymphomas, including *TERT, SOX11, DNMT3A, HVCN1*, and *RCOR2*. The association between hypermethylation and reduced expression of *TERT* in EBV-positive BL was compelling given its association with the reactivation of the EBV lytic cycle(30),(31).

Pathway enrichment analysis of 441 negatively correlated genes revealed an enrichment for RNA polymerase II transcription regulatory region sequence-specific DNA binding (Q = 0.03; **Supplemental Table 6**). Overall, it appears that the hypermethylation landscape of EBV-positive BL methylomes could broadly affect the expression of genes responsible for regulating transcription, thereby broadly reprogramming the transcriptome of BL.

Identification and characterization of BL epitypes

To resolve DNA methylation patterns that cannot be explained by EBV status, we used unsupervised clustering and applied non-negative matrix factorization. The optimal result resolved two patient clusters with distinct methylation patterns among these probes (**Supplemental Figure 2A**,**B**). The genomes in one cluster were predominantly hypermethylated, hereafter referred to as HyperBL with the remaining cases having lower methylation of most probes (HypoBL). Using the cluster assignments as training data, we implemented a statistical model to determine the probability that each BL belonged to either epigenetic subtype (epitype). The optimized classifier utilized 60 probes and had a high accuracy (88%) in the training set of 65% of samples (**Supplemental Table 7**) and similar accuracy (86%) when applied to the held-out validation set. 27 of the training samples (13 HyperBL and 14 HypoBL) were unclassified by this approach due to an intermediate level of methylation at these probes (**Figure 3A, Supplemental Table 8**). These epitype labels were used for all subsequent analyses. We noted an even representation of adult and pediatric cases between HyperBL and HypoBL. Although HypoBL was approximately equally split between EBV-positive (N=56/88) and EBV-negative (N=32/88) tumors, HyperBL had an

enrichment of EBV-positive (N=78/102) compared to EBV-negative tumors (N=24/102, **Figure 3B-C**). When the recently described BL genetic subgroups(9) are considered, DGG-BL was predominantly HyperBL (N=45/70) whereas IC-BL was mostly HypoBL (N=38/65, **Figure 3D**). The methylation differences associated with these epitypes cannot be explained by differences in the tumor sample purity as there was no significant ($P > 0.05$) difference between tumor purities estimated by four separate approaches: Beta values(32), RNA-seq(33), and WGS data using either CNVs (34) or mean VAF of coding mutations (**Supplemental Table 8**).

We next compared the methylation differences between each epitype and normal centroblasts, revealing marked changes in the methylome of HyperBL (**Figure 4A**-**B**). The magnitude of changes in the methylome of HyperBL was greater than the magnitude of changes when EBV-positive BL was compared to normal centroblasts. Specifically, while HyperBL displayed a similar number of hypermethylated DMPs compared to EBV-positive BL (57,596 vs 61,336) they harbored significantly ($P < 0.001$) more hypomethylated DMPs (83,933 vs 47,609). The methylation features of HypoBL more closely resembled that of centroblasts, with fewer DMPs (8227 hypermethylated, 32581 hypomethylated). Directly comparing HyperBL and HypoBL revealed 19832 significant DMPs ($Q < 0.05$ and $|\text{logfoldchange}| > 0.25$). Most of these were methylated in HyperBL and were predominantly found in openSeas (8733), followed by CpG islands (4233), CpG shores (2957), and CpG shelves (1068). The analysis of chromatin and transcriptional state annotations revealed a significant enrichment ($P < 0.001$) of hypermethylated DMPs in promoters (**Figure 4C**). Performing the region level analyses resolved 5550 DMRs, mostly showing hypermethylation in HyperBL (**Figure 4D**, **Supplemental Table 9**). Interestingly, unlike the result when stratifying on EBV status, the DMR associated with HyperBL correspond to regions where methylation has been acquired compared to normal centroblasts and HypoBL. Overall, the hypermethylation patterns in HyperBL were more distinct from those found in normal centroblasts, whereas EBV-positive BL exhibits methylation patterns closer to normal centroblasts. (**Figure 2B**, **Figure 4D**).

As before, we searched for genes with evidence for epigenetic silencing, here focusing on DMRs distinguishing the epitypes. These regions were associated with 2825 genes, of which 210 exhibited significant correlations with expression ($R^2 > 0.4$ and $Q < 0.01$) (**Figure 4E**, **Supplemental Table 10**). These include several genes that are mutated in DLBCL and, to a lesser extent, BL. The hypermethylation and reduced expression of *IRF4* was intriguing, given the role of this transcription factor with terminal differentiation. Within the DMRs, we also observed hypermethylation of several *IRF4* target genes, all of which were negatively correlated with expression in HyperBL. These genes are involved in the NF-kB pathway, important in plasma cell differentiation and are associated with activated B-cell-like (ABC) DLBCL. Hypermethylation of the *TET2* promoter was associated with significantly lower expression ($R^2 = -0.29$, $P < 0.001$). When compared by epitype, HyperBLs exhibited significantly lower *TET2* expression than HypoBL (*P*<0.001; **Supplemental Figure 2C**). Moreover, comparing *TET2* expression across various different B-cell lymphoma subtypes, we found HyperBL to have the lowest *TET2* expression (mean vst = 9.98; **Supplemental Figure 2D**). Given its role in CpG demethylation, it is possible that lower *TET2* expression in HyperBL permits malignant cells to progressively acquire increased levels of methylation.

The methylation patterns in B-cell neoplasms retain some features of the epigenome of their cell-of-origin(18,35–37). We analyzed the BLs along with 1595 methylation profiles from a variety of B-cell neoplasms. By principal component analysis, the BLs were generally separated from other malignancies with the exception of DLBCL. There was also a general separation of HyperBL and HypoBL. Relative to the trajectory of B-cell differentiation, the methylation profiles of HypoBLs were closer to those of normal memory B or plasma cells[15-16] (**Supplemental Figure 2E**). HypoBL also exhibited a closer association with m-CLL and C2.MCL, both of which have proposed origins in memory B cells. The proximity of HypoBL to normal memory B and plasma cell populations, m-CLL and C2.MCL further supports a model wherein the two epitypes arise from distinct cells-of-origin.

To further resolve the underlying factors contributing to HyperBL, we explored whether the affected regions were associated with specific regulatory features. As

transcription factors (TF) can promote hypomethylation of regulatory regions, we performed TF motif discovery within the DMRs[24,25]. This identified motifs associated with transcription factor families responsible for B-cell development, including the NF-kB family, *IRF4*, and *PRDM1* (**Figure 4F, Supplemental Table 11**). In normal B-cell development, increased expression of these transcription factors promotes terminal differentiation and is linked to the hypomethylation of their associated targets in memory B cells and plasma cells(15). These observations imply that these TFs are either active or primed for activation within HypoBL. Additionally, the activation of TFs associated with memory B cells and plasma cells may indicate that the two epitypes arise from B cells at distinct differentiation stages. When this analysis was performed using only DMRs associated with CpG islands, enhancers, and promoters, and not with gene body or intergenic regions, we obtained comparable results (**Supplemental Figure 3A**). For comparison, we conducted similar analyses based on EBV status, which resulted in fewer enrichments for motifs, leaving only the zinc-finger and AP-2 family TFs as significant (**Supplemental Figure 3B, Supplemental Table 12**). The absence of B-cell specific TF binding motifs among the EBV-related DMRs suggests that the hypermethylation in EBV-positive BL is not a consequence of inactivation of their cognate TF but is more plausibly attributable to pressures imposed by the viral infection.

Using epiCMIT[16], a computational approach that infers historical mitotic activity based on specific methylation changes, we estimated the relative number of cell divisions in each tumor. HyperBL samples exhibited significantly higher epiCMIT values compared to HypoBL (*P*=0.001; **Supplemental Figure 3C**). When comparing epiCMIT values across B-cell malignancies, we noted a trend towards higher epiCMIT values in the entities thought to arise from B cells closer to terminal differentiation. This could be explained by the cumulative effect of mitotic cell divisions during normal differentiation and following malignant transformation(18) (**Supplemental Figure 3D**).

To explore methylation changes comprehensively, we sequenced 9 of these samples using long-read sequencing. Nanopolish reported the CpG methylation status of ~28 million CpG sites, representing nearly 99% of all CpGs in the genome. There was strong correlation between these beta values and beta values from the EPIC array

(**Supplemental Figure 4A**). We then performed differential methylation analysis to compare the two epitypes. This revealed a total of 984,369 CpGs that exhibited differential methylation between HypoBL and HyperBL, along with 34,008 DMRs (**Supplemental Figure 4B**). Most of these regions displayed hypermethylation in HyperBLs and they were prominently enriched in promoter regions. These DMRs were strongly correlated with those identified through the EPIC array, providing additional support for our findings (Spearman correlation, *P* value < 0.001).

The long-read sequencing allows the resolution of CpG methylation status of each DNA molecule, which yields a digital measurement. This affords the opportunity to quantify the heterogeneity of methylation across DNA molecules (entropy)(38) that is not possible with analog measurements such as beta values. We compared the entropy at CpG sites within genes and regulatory elements in both epitypes. The methylation entropy of promoter regions was substantially higher in HyperBL compared to HypoBL, indicating a more uniform DNA methylation state at these sites in HypoBL (**Supplemental Figure 4C**,**D**). This may imply that the hypomethylation of DMRs in promoter regions of HypoBL results from strong selective pressure, involving active enzymatic DNA demethylation. Using an overrepresentation analysis of promoter regions with low entropy in HypoBL, we found an enrichment of genes involved in hematopoietic differentiation (GO:0048534) and hematopoietic stem cell differentiation (GO:0060218; **Supplemental Figure 4E**). The low entropy within these promoters suggest their epigenetic de-repression or sustained maintenance of de-repression in HypoBL. These genes are activated in long-lived memory B cells, which are considered to possess a stemness reminiscent of CD8+ memory T cells(39–41). These results further indicate a distinct cell-of-origin for HypoBL.

Genetic features of epitypes

Utilizing previously published genome-wide mutation profiles(9), we next compared the epitypes for differences in mutation patterns (**Figure 5A**) and mutation burden. The latter was significantly higher in HyperBL (*P*<0.001; **Figure 5B**). This disparity was consistent when restricted to coding mutations (*P*<0.001) or known BL

driver mutations (*P*=0.012). Comparing the mutation frequency in each BL gene between epitypes and by modules of genes with related function **(Supplemental Figure 5A)** revealed several distinctions. At the single-gene level, only *ID3* exhibited a higher frequency of mutations in HypoBL. For modules, genes associated with epigenetic regulation and antigen presentation by major histocompatibility complex (MHC) exhibited a higher frequency of mutations in HyperBL (***Q* <0.01, **Q* <0.05, Fisher's Exact) (**Supplemental Figure 5A**). Interestingly, *ID3* promoter methylation was significantly higher in patients lacking *ID3* mutations, implying epigenetic silencing may be an alternative avenue to reducing ID3 function in BL (*P*=0.005, Fisher's Exact).

Considering the overlap between HyperBL with EBV-positivity, we searched for differences within each epitype by stratifying patients into four categories using both epitype and EBV status. Interestingly, whereas *TP53* mutations were common in EBV-positive HyperBL (HyperBL+)(24%) they are absent among the EBV-positive HypoBL (HypoBL-) (**Figure 5A**). Among EBV- cases, HyperBL- had mutations in at least one epigenetic modifier, whereas no mutations in these were found in HypoBL-. For example, *SMARCA4* mutations were common in EBV-negative BL overall but exclusive to HyperBL- (**Figure 5A**). This interplay implies distinct selective paths to lymphomagenesis that may be influenced by both EBV status and epitype.

The genome-wide landscape of mutations in HyperBL was strikingly different, with higher global mutation burden along with more coding mutations and mutations in BL genes (**Figure 5B**). This contrasts with prior observations based on EBV status in BL, where EBV-positive BLs possess a higher frequency of global and coding mutations but notably fewer driver mutations. Using mutational signatures, we compared the single base substitution exposures between the epitypes. Each of SBS1 (clock-like), SBS5 (clock-like/age-associated), and SBS9 (associated with *AICDA* and polymerase η activity) exhibited significant differences between the epitypes (**Figure 5C**). The mutation levels attributed to all three signatures were consistently higher in HyperBL (*P*<0.001). Considering the previous link between the SBS9 mutation signature and EBV-positive BL, we considered whether EBV status or epitype was the primary factor influencing mutation rates. Using a linear model, we found that mutation rates linked to

SBS9, SBS1 and SBS5 were more significantly associated with epitype than EBV status. Similarly, epitype more accurately explained the variations in mutation burden, encompassing global, coding, and driver mutations. This underscores the unique molecular attributes linked to each BL epitype, suggesting that this classification is more relevant than EBV status when considering the molecular features of the BL genome.

Mutations occurring within regions known for aSHM are attributed to the abnormal activity of activation-induced cytidine deaminase (AID), a pattern that has been observed predominantly in EBV-positive BLs. By comparing the level of aSHM-associated mutations between the two epitypes, a higher rate was found in the HyperBL genomes (**Supplemental Figure 5B**). We assessed whether aSHM mutations at each of these sites, and the total burden of aSHM-associated mutations, were more strongly associated with epitype or EBV status. Again, epitype membership consistently exhibited a stronger association than EBV status (**Supplemental Figure 5B, Supplemental Table 13**). While this does not exclude that aSHM are associated with EBV status, they are more strongly correlated with epitype membership regardless of viral infection.

Additional molecular distinctions between the epitypes

Although DNA methylation influences expression, it was unclear whether the epitypes would harbor consistent gene expression changes shaped by their unique DNA methylation landscape. Comparing the epitypes directly, while controlling for EBV status, revealed only 218 genes with significant differential expression (padj < 0.01 and |log2FoldChange| > 1), including *TET2, IRF4, CD44*, and *CD24* (**Figure 6A, Supplemental Table 14**). The differential expression of *CD44, CD24* and *IRF4* are, respectively, of interest due to their dynamic expression in the GC reaction (CD44 and CD24) and role in plasmablastic differentiation(23,42,43). Higher *IRF4* expression in HypoBL is reminiscent with the difference we observed between IC-BL and DGG-BL genetic subtypes. Although the variation in *IRF4* expression was more pronounced when categorized by genetic subtype as opposed to epitype, the parallels in the distribution of *IRF4* expression based on both criteria implies that these classification

methods are complementary yet capture distinct biological features (**Supplemental Figure 5C,D**).

We analyzed the RNA-seq data to infer B-cell receptor expression and search for evidence of class-switch recombination (CSR). We classified each sample based on IGH constant gene expression (*IGHM, IGHD, IGHA, IGHG*, or *IGHE*) (**Supplemental Figure 5E**) and observed significantly higher (P = 0.03) representation of IGHG among the HyperBLs. To complement this, we attributed the mechanism underlying each of the *MYC* rearrangements. Whereas 32% of HyperBL cases had a *MYC* translocation that could be attributed to CSR, the majority (56%) of *MYC* translocations in HypoBL could be attributed to CSR (**Supplemental Figure 5F**). Considering the higher rate of IGHM expression in HypoBL and the significantly higher (P = 0.003) proportion of *MYC* rearrangements resulting from CSR, we conclude that HypoBL may derive from memory B cell that has attempted, and failed, to undergo CSR(44–47).

Given the limited number of genes with consistent expression differences, we used gene set enrichment analyses to search for more subtle effects on genes with related function. We identified 32 differentially expressed gene sets(23), including two associated with the *IRF4* network (**Figure 6B, Supplemental Figure 6, Supplemental Table 15**). HypoBL displayed elevated expression of genes involved in *IRF4* induction in ABC-DLBCL, other ABC-DLBCL-related pathways, as well as a memory B cell precursor pathway (**Figure 6B**). HyperBL showed elevated expression of genes in genes related to *IRF4* repression in GC B-cell-like (GCB)-DLBCL and other pathways associated with GCB-DLBCL (**Figure 6B**). The elevated pathways in each epitype is consistent with earlier results that suggest that HypoBL originate from a memory B cell whereas HyperBL likely originates from centroblasts.

Relationship between epitypes and patient outcomes

We conducted Kaplan-Meier survival analyses within the adult cases, for whom the follow-up data was most complete. Among these cases, patients with HyperBL had significantly shorter progression-free and overall survival. This difference remains

significant (P = 0.035 and P = 0.014, respectively) when the unclassified patients are also included in the analysis (**Figure 7A-B**). Moreover, the difference cannot be explained by other molecular biomarkers such as *TP53* mutation status(9) because frequencies of *TP53* mutations were relatively balanced in both epitypes within the adult patients and epitype status remained significant in a multivariate analysis with *TP53* mutation status (**Supplemental Figure 7A-D**). Although this finding requires confirmation in additional cohorts, the potential association between epitype and outcomes highlights DNA methylation status as a potential new prognostic biomarker for BL.

Discussion

Previous studies provided limited exploration of the BL methylome and its contribution to pathogenesis, primarily due to limited sample sizes and the narrow scope of comparisons focusing solely on EBV status or comparisons to follicular lymphoma(20,48). Here, we aimed to address these limitations by conducting a comprehensive examination of the BL methylome. Our findings confirm the influence of EBV status on DNA methylation and provide further evidence of the connection between EBV tumor positivity and hypermethylation within the DNA methylome, particularly in enhancer regions. Moreover, our investigation of the most variable CpGs in EBV-positive and EBV-negative BL allowed us to identify two distinct epitypes that extend beyond the classification based on EBV status alone.

EBV infection is detected in a variable proportion of BL tumors, differing by clinical variant and age. EBV-positive BL cases exhibit distinct characteristics, including a lower number of driver mutations specifically related to apoptotic genes, higher rates of aSHM, and increased AICDA activity(8,9). Consistent with a previous study(8), we find that EBV-positive BLs display significant hypermethylation patterns in their DNA methylomes compared to EBV-negative BLs and normal centroblast samples. This hypermethylation predominantly occurs within OpenSea annotations, targets a number of genes previously identified as hypermethylated across B-cell lymphomas, and shows enrichment in enhancer regions(49–52). Our results are consistent with previous studies

showing that EBV can promote DNA hypermethylation aimed at precisely controlling the expression of specific genes, such as *PRDM1*, which can trigger EBV lytic reactivation, thereby increasing its capacity to evade detection by the immune system(30,49,50,52).

The occurrence of focal hypermethylation within BL has been described in the context of EBV(49,50,53,54). While HyperBL is dominated by EBV-positive samples, HyperBL occurrence in EBV-negative cases suggests EBV infection may not be a prerequisite for this phenomenon. *TET2* hypermethylation in HyperBL was recurrent and associated with reduced expression. Further, we note the highest incidence of mutations affecting regulators of epigenetic state among the genomes of HyperBLs. Indeed, among the 21 EBV- HyperBL cases, 3 had loss-of-function mutations affecting *TET2* whereas there were none observed among the 63 EBV+ HyperBLs (P= 0.01396, Fisher's Exact). Given the suggested role of *TET2* mutations in genome hypermethylation in DLBCL(55,56), it is plausible that the results of *TET2* mutation or hypermutation along with mutations in other epigenetic-related genes contribute to the hypermethylation pattern in the DNA methylome applicable to HyperBL, particularly in the absence of EBV.

HyperBL also has elevated mutation frequencies in genes associated with epigenetic regulation and antigen presentation (**Supplemental Figure 5A**). In contrast, HypoBL exhibited a higher prevalence in genes associated with, notably, *ID3* mutations. An inverse correlation emerged between *ID3* mutation status and promoter methylation patterns. Specifically, we observed a significant prevalence of *ID3* promoter hypermethylation in patients devoid of *ID3* mutations. This pattern suggests that the inactivation of *ID3* can occur either through genetic mutations or through epigenetic modifications, such as promoter hypermethylation, underscoring the multifaceted mechanisms that contribute to the functional loss of this gene. While we noted a pronounced correlation between HyperBL and EBV-positive BL, distinct variations exist sufficient to identify epitype as biologically distinct from EBV status. For example, approximately a quarter of EBV-positive HyperBL cases exhibited *TP53* mutations, but none were observed in EBV-positive HypoBL. In the context of EBV-negative BL, all HyperBL cases exhibited mutations in at least one gene related to epigenetic regulation;

conversely, those mutations were not observed in EBV-negative HypoBL. These observed disparities underscore a multifaceted interaction and unique selective pressures that characterize the process of lymphomagenesis in each epitype.

Non-coding mutations consistent with aSHM is a pattern previously linked to EBV-positive status in BLs(8,9). However, our findings reveal a notable enrichment of aSHM in HyperBL, a pattern that is correlated with but not wholly explained by EBV-positive BL status per se. Our comparative analysis established a stronger linkage of aSHM rates with epitype after controlling for EBV status. Based on this comprehensive analysis, we infer that both HyperBL and HypoBL encompass EBV-positive and EBV-negative tumors, and that each epitype represents a unique biological entity, independent of EBV status. Our data underscore that complexity of mechanisms leading variations in aSHM, mutation burden, and mutation signatures are influenced by EBV status, and more specifically the epitype. Thus, epitype may serve as a more relevant biomarker of mutational processes within BL genomes than EBV status, offering a new way to study BL pathogenesis.

The use of gene expression for classification has been instrumental in understanding the diverse nature and prognosis of various NHLs, including BL. In DLBCL, it has offered insights into distinct biological underpinnings and facilitated its categorization into different COO subgroups(57–59). However, when extrapolated to other lymphomas, this approach has shown limitations particularly in diseases like chronic lymphocytic leukemia (CLL) and mantle cell lymphoma (MCL), where distinguishing COO through gene expression profiling has been challenging(16,17). In our study, we describe epitypes within BL that potentially illuminate COO hypothesis, each associated with distinct survival outcomes. HyperBL and HypoBL are distinguished by notable molecular characteristics, with biological and transcriptomic distinctions that echo the COO subgroups in DLBCL. Specifically, HyperBL was associated with pronounced promoter hypermethylation and inferior overall survival, appearing epigenetically closer to the GC B cell. Conversely, HypoBL appears epigenetically closer to a memory B cell, maintaining methylation patterns typical of normal B cells at this developmental phase and is associated with better survival

outcomes. Furthermore, we found reduced entropy in the promoters of genes tied to hematopoietic stem cell differentiation in HypoBL.

In our observations, HypoBL exhibited a pronounced elevation in *IRF4* expression compared to HyperBL, a trend that aligns with the distinctions noted between IC-BL and DGG-BL when sorted by genetic subtypes. While the disparity in *IRF4* expression is more accentuated under genetic subtype classification, the similarities in *IRF4* expression distribution under both genetic and epitype categorizations suggest these two classification approaches, though capturing different nuances, are complementary. In HypoBL, there is an amplified expression of genes associated with *IRF4* induction in ABC DLBCL and a pathway indicative of a memory B-cell precursors. These findings, coupled with distinct methylation patterns, suggest the possibility of unique cellular origins for HyperBL and HypoBL, each carrying distinct biological and clinical connotations.

We have identified a subgroup within BL characterized by elevated levels of DNA methylation and have also shown elevated levels of methylation present within EBV-positive BL. The shared occurrence of hypermethylation in both EBV-positive BL and the HyperBL subgroup highlights the potential value of investigating demethylating agents as part of BL treatment approaches. Further detailed investigations are essential to confirm these hypotheses and understand their broader implications in the context of BL pathogenesis and treatment.

Methods

Case accrual

This study was performed in accordance with the ethical standards of the Declaration of Helsinki. All contributing sites provided Institutional Review Board approvals for the use of tissues submitted for molecular characterization. Written informed consent was obtained from the parents or guardians of the children and written

informed assent was obtained from children aged seven years or older prior to enrolment.

Methylation data collection and integration

Samples were analyzed by Illumina Infinium™ MethylationEPIC BeadChip arrays following the manufacturer's protocol. Beginning with a compendium of 866,836 probes, a systematic filtration process was executed employing the minfi R package(60). Initial exclusion criteria encompassed the removal of 977 CpG probes with inadequate signal detection, followed by the omission of 30,435 CpGs corresponding to single nucleotide polymorphisms (SNPs) and 19,298 CpGs located on sex chromosomes. The retained subset of 812,126 CpGs exhibited a detection p-value <= 0.05 in over 10% of samples. Subsequent exclusions targeted samples with suboptimal intensity signals or problematic probe conversions, culminating in the elimination of 10 samples. Post-filtration, a total of 218 BL samples and 6 normal centroblast samples(8) were preserved for further analysis, encompassing DNA methylation values for 812,126 CpG probes. These values were normalized employing the Subset-quantile Within Array Normalization (SWAN) algorithm. Comprehensive annotation of all CpGs was facilitated by the IlluminaHumanMethylationEPICanno.ilm10b4.hg19 (version 0.6) R package. To enable comparative analyses across B-cell neoplasms and normal B cell types, we analyzed methylation data from 1595 B cell samples from Duran-Ferrer *et al*(18) and integrated these with our data. Subsequent analyses were conducted using the 452,679 CpG sites common to both methylation array platforms.

Differential DNA methylation analyses

DMRcate(61) was used to identify differentially methylated probes (DMPs) and differentially methylated regions (DMRs) using pre-processed beta values as input. DMPs required absolute log fold change (logFC) greater than 0.25 and adjusted p-value < 0.05. DMRs required a mean methylation difference >0.1, a false discovery rate (FDR) < 0.05, and at least 4 CpG sites. The analysis of differential methylation between

epitypes included EBV status as a variable in the design model to facilitate the identification of DMPs and DMRs not directly associated with EBV status.

To quantify enrichment among DMPs, the entire complement of CpGs from the EPIC array was utilized to establish a background reference. To ascertain enrichments of CpG probes within regulatory regions (ChromHMM), we assessed the proportion of differentially methylated CpGs located in these regions relative to the proportion of background EPIC array probes within the same regulatory contexts. Enrichment significance was determined through the calculation of odds ratio scores using Fisher's Exact ($P < 0.01$).

NMF clustering

Non-negative Matrix Factorization (NMF) clustering relied on the NMF package (0.23.0). 5000 probes with the highest variance within each of the EBV+ and EBV- samples were used as input (10,000 probes total). To ascertain the optimal factorization rank for NMF, 100 bootstrapped iterations ranging from rank 2 to 5 were executed, with the optimal rank identified through the assessment of the cophenetic coefficient, dispersion, and silhouette metrics. Upon determining a factorization rank of 2 as the best fit, various algorithms, specifically those of Lee, Brunet, offset, and nsNMF, were evaluated by iteratively running them on the feature matrix until convergence was achieved. The Brunet algorithm emerged as the most suitable for subsequent analysis based on these evaluations. Throughout the estimation of factorization rank, algorithm selection, and NMF computations, a consistent random seed value of 12345 was employed to ensure reproducibility. Sample clustering was determined using the Euclidean distance metric, with individual sample subgroup membership ascertained by extracting matrix coefficients and assigning samples to the subgroup corresponding to their highest coefficient.

Training BL epitype classifier

Using the LPS R package, a linear predictor model was engineered to ascertain the optimal probe count necessary for categorizing samples into HyperBL and HypoBL epigenetic subtypes, similar to that described in Wright *et al.*(57) Initially, a training set composed of 77 HyperBL and 66 HypoBL specimens was compiled, in conjunction with a validation set containing 40 HyperBL and 35 HypoBL samples. A series of models, each incorporating a variable number of probes, underwent assessment within the training set, employing a leave-one-out cross-validation strategy to identify the optimal number of probes for the LPS model. The iteration incorporating 60 probes, selected from an initial pool of 10,000 probes employed for NMF clustering, demonstrated the lowest mean error rate. When this model was applied to the validation cohort, the LPS score distribution within each epigenetic subtype mirrored that of the training cohort, effectively negating the presence of model overfitting. Employing probability determinations based on Bayes' theorem, a 90% certainty threshold was established for definitive epitype classification. Accordingly, samples were designated to an epitype only if the likelihood of their belonging met or exceeded the 90% threshold; otherwise, they were considered unclassified.

Genome-wide methylation profiling with long reads

The occurrence of 5mC in the long-read sequencing data was determined with Nanopolish(62) (v. 0.8.5) using the combination of raw signals from the sequencer found in the fast5 files and aligned PromethION bam files as input. Methylation frequencies (beta values) at reference CpGs were then summarized using the Nanopolish “calculate methylation frequency” script and enabling the split flag (-s) to obtain single CpG methylation. Beta values where subset down to retain only the autosomes (chr:1-22), removing the sex (X & Y) chromosomes. The analysis of long-read methylation profiling was done using LCR-modules and the code is available at https://github.com/LCR-BCCRC/lcr-modules.

The DSS(63) R package was utilized to execute differential methylation analyses on PromethION sequencing data, adhering to the procedure outlined in the official tutorial. Differentially methylated CpG sites were characterized by a mean methylation differential exceeding 0.1 between epitypes, coupled with a false discovery rate (FDR) < 0.05. Differentially methylated regions (DMRs) were identified under the criteria of a mean methylation difference surpassing 0.1 between epitypes, an FDR < 0.05, and a composition of at least 4 CpGs.

Estimating methylation entropy from long-read data

Shannon entropy scores were computed from methylation values from Nanopolish using CpelNano(64). Comparative analysis of these scores across varied transcriptional states between the two epitypes was conducted utilizing the Wilcoxon rank-sum test (base R, RRID:SCR_001905). Promoter regions exhibiting significant (Q < 0.05) differences in entropy scores, specifically those promoters hypomethylated with low entropy scores in HypoBL were further analyzed for enrichment using the rGREAT(65), package within R following the recommended guidelines outlined in the tutorial. The resulting Gene Ontology (GO) Biological Processes enrichment data were compiled via the getEnrichmentTables function to obtain the final over-enrichment results table.

DNA Extraction and PromethION sequencing

High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA Kit (cat. no. 67563, QIAGEN, Germantown, MD, USA). Genomic libraries were prepared, conforming to Oxford Nanopore Technologies' protocols, using the SQK-LSK109 Ligation Library Kit. NEB Ultra II kit (New England Biolabs, Ipswich, MA, USA, cat. no. E7646A) was used for end-repair and A-tailing. NEBNext quick ligase (E6056S) was used to ligate the Oxford Nanopore adapters. A final size selection of 0.4:1 ratio (magnetic beads to library) was done to select against smaller molecules. PromethION sequencing proceeded using the R9.4.1 pore flow cell on the PromethION Alpha-Beta instrument and the beta release software version 19.06.9 (MinKNOW

v3.4.6, GUI v3.4.12). A DNase I nuclease flush (Invitrogen cat no. AM2222) was performed after 18 hours as per Oxford Nanopore protocol, version NFL_9076_v109_revD_08Oct2048. Base calling from the resulting fast5 files was performed with guppy v3.2.1 and the reads were mapped to GRCh38 using minimap2.

Annotation of probes with chromatin and transcriptional state

To refine our annotations, probes were mapped to chromatin and transcriptional states, as defined by the lymphoblastoid B-cell line GM12878 obtained from the ChromHMM track of the UCSC Genome Browser. The chromatin states were partitioned as follows: states 1–3 were classified as Promoter regions, indicative of active, weak, and poised promoters; states 4–7 were designated as Enhancer regions, encompassing strong and weak enhancers; state 8 delineated Insulator regions; state 9 was attributed to Transcriptional transition; state 10 to Transcriptional elongation; state 11 to Weak transcription; state 12 to Polycomb-repressed regions; and state 13 to Heterochromatin.

Transcription factor binding motif enrichment analyses

JASPAR enrichment analysis tool was employed to perform transcription factor binding motif enrichment analysis. Input for the analysis consisted of differentially methylated regions (DMRs) delineated by EBV status and epigenetic subtype. These regions were juxtaposed against a carefully constructed background comprised from EPIC array probes. Prior to analysis with JASPAR, both DMRs and background data were converted from hg19 to hg38 genomic coordinates. Adhering to established JASPAR protocols, enrichment calculations were conducted within a specified universe of genomic regions. The DMRs constituted the user-defined input set, while the background constructed from the EPIC array probes served as the user background/universe set.

Gene expression and pathway analyses

Gene expression was quantified at the transcript level using Salmon as described previously in Thomas *et al*.(9) Differential gene expression analysis was performed using edgeR between the epitypes, where RNA-seq data was available. The

experimental design model employed to determine differentially expressed genes between epitypes allowed for the controlling of variables including cell sorting status, patient sex, and EBV status, while identifying differentially expressed genes based on epitype. Cut-offs of padj < 0.01 and |log2FoldChange| > 1.5 were employed to determine significant differentially expressed genes. Differentially expressed genes were further visualized using R package ComplexHeatmap (version 2.2.0). Due to the disproportionate composition of EBV-positive samples between epitypes, and to ensure the most significant differentially expressed genes were best explained by epitypes, we created a linear model for said genes. The model included both EBV status and epitype as variables.

Gene set enrichment analysis was performed using GSVA (version 1.34.0) comparing the two epitypes. Normalized expression data and gene sets obtained from signatureDB were used as input for gene set variation analyses. Gene sets were filtered for a minimum of 5 genes and a maximum of 500. Significant differentially enriched gene sets were determined based on a cut-off of p.adj < 0.01 and visualized using the R package ComplexHeatmap.

Mitotic clock estimates

Utilizing the EPIC array dataset, methylation-based mitotic clock estimates were ascertained via the epiCMIT (https://github.com/Duran-FerrerM/Pan-B-cell-methylome), in adherence to the protocol specified in the accompanying example documentation. Concurrently, mitotic clock estimates were obtained from a cohort of 1595 B-cell samples, as detailed in Duran-Ferrer *et al*.(18) These estimates facilitated subsequent comparative analyses across various B-cell malignancies and against baseline measurements from normal B-cell populations comparing the proliferative histories of the various entities.

Data Availability Statement

All primary whole genome, long-read promethION, and transcriptome sequencing data, as well as clinical data used in this publication can be found on the National

Cancer Institute's Genome Data Commons that can be accessed directly at https://portal.gdc.cancer.gov/projects/CGCI-BLGSP. The raw DNA methylation profiling data analyzed with EPIC arrays is deposited and publicly available through Gene Expression Omnibus (GEO) under accession number GSE292690 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE292690). All custom bioinformatics workflows, scripts, postprocessing, and visualization functions are openly available on GitHub through repository Lymphoid Cancer Research (LCR) modules (https://github.com/LCR-BCCRC/lcr-modules), LCR scripts (https://github.com/LCR-BCCRC/lcr-scripts), and GAMBLR.open package (https://github.com/morinlab/GAMBLR.open).

Acknowledgements

The authors thank the Foundation for Burkitt Lymphoma Research Working Group for interesting discussions. The authors also acknowledge the Information Management Systems (Silver Spring, MD), Westat, Inc (Rockville, MD), and African Field Epidemiology Network (Kampala, Uganda) for coordinating The Epidemiology of Burkitt lymphoma in East African children and minors (EMBLEM) study fieldwork in Uganda. The authors are grateful for contributions from various groups at Canada's Michael Smith Genome Sciences Centre, including those from the Biospecimen, Library Construction, Sequencing, Bioinformatics, Technology Development, Quality Assurance, Laboratory Information Management System, Purchasing, and Project Management teams. The authors also thank The Biorepository of St. Jude Children's Research Hospital (National Cancer Institute grants P30 CA021765 and R35 CA197695 to C.G.M.).

This work has been funded in part by the Foundation for Burkitt Lymphoma Research (http://www.foundationforburkittlymphoma.org) and in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health (NIH), under contract no. 75N91019D00024, task order no. 75N91020F00003, contract no. HHSN261200800001E, contract no. HHSN261201100063C, and contract no. HHSN261201100007I (Division of Cancer Epidemiology and Genetics), and in part (S.J.R.) by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, NIH. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government. This work was supported by a Terry Fox New Investigator Award (No. 1043) and by an operating grant from the Canadian Institutes for Health Research and a New Investigator Award from the Canadian Institutes for Health Research (R.D.M.). R.D.M. is a Michael Smith Foundation for Health Research Scholar, and D.W.S. is a Michael Smith Foundation for Health Research Health Professional-Investigator. M.A.M. is the recipient of the Canada Research Chair in Genome Science.

M.E. is supported by the CERCA Programme / Generalitat de Catalunya (2021 SGR01494), Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033/ERDF ‘A way to make Europe’ (PID2021-125282OB-I00) and the Cellex Foundation (CEL007).
This article is dedicated to the memory of Daniela S. Gerhard.

Authorship Contributions

N.T. and K.D. analyzed the data, produced the figures and tables, and with R.D.M. wrote the manuscript, J. Bethony, C.C., T.G., N.L.H., E.S.J., S.M.M., C.G.M., A.J.M., A.N., M.A.M., D.W.S., and M.E. ;B.M.G., L.K.H., M.E., and C.A.G.-P. helped with data generation and analyses; M.A.D., N.B.G., and H.P. managed the project and coordinated data deposition; A.C.B., J. Bowen, C.C., J.M.G.-F., T.G.G., F.E.L., S.M.M., C.G.M., C.N., A.N., M.D.O., J.O., G.O., D.W.S., and M.N. contributed samples to the study; J. Bowen, J.M.G.-F., H.P., and A.S.G. collected sample metadata from tissue source sites; T.G., N.L.H., E.S.J., and S.H.S. performed consensus pathology review; C.C., T.G.G., and E.S.J. reviewed and advised on consensus anatomic site classification; J.D.I., R.D.M., and L.M.S. designed the study; M.A.M., R.D.M., and L.M.S. directed the study; all authors contributed to the interpretation of the data, reviewed the manuscript, and approved it for submission.

Disclosure of Conflicts of Interest

R.D.M. and D.W.S. are named inventors on a patent application describing the double-hit signature. C.G.M. received research funding from Pfizer, AbbVie; advisory board member at Illumina and speakers bureau at Amgen. A.N. received research funding from Pharmacyclics/AbbVie, Kite/Gilead, and Cornerstone; was a consultant for Janssen, Morphosys, Cornerstone, Epizyme, EUSA Pharma, TG Therapeutics, ADC Therapeutics, and Astra Zeneca; and has received honoraria from Pharmacyclics/AbbVie. M.E. reports grants from Incyte and personal fees from

Quimatryx outside the submitted work. All other authors declare no competing financial and/or non-financial interests in relation to the work described.

References

1. Hämmerl L, Colombet M, Rochford R, Ogwang DM, Parkin DM. The burden of Burkitt lymphoma in Africa. Infectious Agents and Cancer. 2019;14:17.

2. Stefan DC, Lutchman R. Burkitt lymphoma: epidemiological features and survival in a South African centre. Infectious Agents and Cancer. 2014;9:19.

3. Bouska A, Bi C, Lone W, Zhang W, Kedwaii A, Heavican T, et al. Adult high-grade B-cell lymphoma with Burkitt lymphoma signature: genomic features and potential therapeutic targets. Blood. 2017;130:1819–31.

4. Schmitz R, Ceribelli M, Pittaluga S, Wright G, Staudt LM. Oncogenic Mechanisms in Burkitt Lymphoma. Cold Spring Harb Perspect Med. 2014;4:a014282.

5. Love C, Sun Z, Jima D, Li G, Zhang J, Miles R, et al. The genetic landscape of mutations in Burkitt lymphoma. Nat Genet. Nature Publishing Group; 2012;44:1321–5.

6. Molyneux EM, Rochford R, Griffin B, Newton R, Jackson G, Menon G, et al. Burkitt's lymphoma. The Lancet. 2012;379:1234–44.

7. Magrath I. Epidemiology: clues to the pathogenesis of Burkitt lymphoma. British Journal of Haematology. 2012;156:744–56.

8. Grande BM, Gerhard DS, Jiang A, Griner NB, Abramson JS, Alexander TB, et al. Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. Blood. 2019;133:1313–24.

9. Thomas N, Dreval K, Gerhard DS, Hilton LK, Abramson JS, Ambinder RF, et al. Genetic subgroups inform on pathobiology in adult and pediatric Burkitt lymphoma. Blood. 2023;141:904–16.

10. Alaggio R, Amador C, Anagnostopoulos I, Attygalle AD, Araujo IB de O, Berti E, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. Leukemia. Nature Publishing Group; 2022;36:1720–48.

11. Bergman Y, Cedar H. DNA methylation dynamics in health and disease. Nat Struct Mol Biol. Nature Publishing Group; 2013;20:274–81.

12. Esteller M. Epigenetic gene silencing in cancer: the DNA hypermethylome. Human Molecular Genetics. 2007;16:R50–9.

13. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, et al. A DNA methylation fingerprint of 1628 human samples. Genome Res. 2012;22:407–19.

14. Xia D, Leon AJ, Yan J, Silva A, Bakhtiari M, Tremblay-LeMay R, et al. DNA Methylation-Based Classification of Small B-Cell Lymphomas: A Proof-of-Principle Study. The Journal of Molecular Diagnostics. Elsevier; 2021;23:1774–86.

15. Kulis M, Merkel A, Heath S, Queirós AC, Schuyler RP, Castellano G, et al. Whole-genome fingerprint of the DNA methylome during human B-cell differentiation. Nat Genet. 2015;47:746–56.

16. Kulis M, Heath S, Bibikova M, Queirós AC, Navarro A, Clot G, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet. Nature Publishing Group; 2012;44:1236–42.

17. Queirós AC, Beekman R, Vilarrasa-Blasi R, Duran-Ferrer M, Clot G, Merkel A, et al. Decoding the DNA methylome of mantle cell lymphoma in the light of the entire B cell lineage. Cancer Cell. 2016;30:806–21.

18. Duran-Ferrer M, Clot G, Nadeu F, Beekman R, Baumann T, Nordlund J, et al. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. Nat Cancer. 2020;1:1066–81.

19. Hummel M, Bentink S, Berger H, Klapper W, Wessendorf S, Barth TFE, et al. A Biologic Definition of Burkitt's Lymphoma from Transcriptional and Genomic Profiling. N Engl J Med. Massachusetts Medical Society; 2006;354:2419–30.

20. Kretzmer H, Bernhart SH, Wang W, Haake A, Weniger MA, Bergmann AK, et al. DNA-methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. Nature genetics. Europe PMC Funders; 2015;47:1316.

21. Shawky SA, El-Borai MH, Khaled HM, Guda I, Mohanad M, Abdellateif MS, et al. The prognostic impact of hypermethylation for a panel of tumor suppressor genes and cell of origin subtype on diffuse large B-cell lymphoma. Mol Biol Rep. 2019;46:4063–76.

22. Hayslip J, Montero A. Tumor suppressor gene methylation in follicular lymphoma: a comprehensive review. Molecular Cancer. 2006;5:44.

23. Victora GD, Dominguez-Sola D, Holmes AB, Deroubaix S, Dalla-Favera R, Nussenzweig MC. Identification of human germinal center light and dark zone cells and their relationship to human B-cell lymphomas. Blood. 2012;120:2240–8.

24. Esteller M, Guo M, Moreno V, Peinado MA, Capella G, Galm O, et al. Hypermethylation-associated Inactivation of the Cellular Retinol-Binding-Protein 1 Gene in Human Cancer1. Cancer Research. 2002;62:5902–5.

25. Esteller M. Profiling aberrant DNA methylation in hematologic neoplasms: a view from the tip of the iceberg. Clinical Immunology. 2003;109:80–8.

26. Amara K, Trimeche M, Ziadi S, Laatiri A, Hachana M, Korbi S. Prognostic significance of aberrant promoter hypermethylation of CpG islands in patients with diffuse large B-cell lymphomas. Annals of Oncology. Elsevier; 2008;19:1774–86.

27. Mohamed G, Talima S, Li L, Wei W, Rudzki Z, Allam RM, et al. Low Expression and Promoter Hypermethylation of the Tumour Suppressor SLIT2, are Associated with Adverse Patient Outcomes in Diffuse Large B Cell Lymphoma. Pathol Oncol Res. 2019;25:1223–31.

28. Grønbæk K, Ralfkiaer U, Dahl C, Hother C, Burns JS, Kassem M, et al. Frequent hypermethylation of DBC1 in malignant lymphoproliferative neoplasms. Modern Pathology. Elsevier; 2008;21:632–8.

29. Coyle KM, Dreval K, Hodson DJ, Morin RD. Audit of B-cell cancer genes. Blood Advances. 2025;9:2019–31.

30. Dolcetti R, Giunco S, Dal Col J, Celeghin A, Mastorci K, De Rossi A. Epstein-Barr virus and telomerase: from cell immortalization to therapy. Infect Agent Cancer. 2014;9:8.

31. Bellon M, Nicot C. Regulation of Telomerase and Telomeres: Human Tumor Viruses Take Control. JNCI: Journal of the National Cancer Institute. 2008;100:98–108.

32. Qin Y, Feng H, Chen M, Wu H, Zheng X. InfiniumPurify: An R package for estimating and accounting for tumor purity in cancer methylation research. Genes & Diseases. 2018;5:43–5.

33. Steen CB, Liu CL, Alizadeh AA, Newman AM. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. Methods Mol Biol. 2020;2117:135–57.

34. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell. 2012;149:994–1007.

35. Gravemeyer J, Spassova I, Verhaegen ME, Dlugosz AA, Hoffmann D, Lange A, et al. DNA-methylation patterns imply a common cellular origin of virus- and UV-associated Merkel cell carcinoma. Oncogene. Nature Publishing Group; 2022;41:37–45.

36. Oakes CC, Martin-Subero JI. Insight into origins, mechanisms, and utility of DNA methylation in B-cell malignancies. Blood. 2018;132:999–1006.

37. Pan H, Jiang Y, Boi M, Tabbò F, Redmond D, Nie K, et al. Epigenomic evolution in diffuse large B-cell lymphomas. Nat Commun. Nature Publishing Group; 2015;6:6921.

38. Xie H, Wang M, de Andrade A, Bonaldo M de F, Galat V, Arndt K, et al. Genome-wide quantitative assessment of variation in DNA methylation patterns. Nucleic Acids Research. 2011;39:4099–108.

39. Kurosaki T, Kometani K, Ise W. Memory B cells. Nat Rev Immunol. Nature Publishing Group; 2015;15:149–59.

40. Luckey CJ, Bhattacharya D, Goldrath AW, Weissman IL, Benoist C, Mathis D. Memory T and memory B cells share a transcriptional program of self-renewal with long-term hematopoietic stem cells. Proc Natl Acad Sci U S A. 2006;103:3304–9.

41. Good-Jacobson KL. Strength in diversity: Phenotypic, functional, and molecular heterogeneity within the memory B cell repertoire. Immunological Reviews. 2018;284:67–78.

42. Ochiai K, Maienschein-Cline M, Simonetti G, Chen J, Rosenthal R, Brink R, et al. Transcriptional Regulation of Germinal Center B and Plasma Cell Fates by Dynamical Control of IRF4. Immunity. 2013;38:918–29.

43. Laidlaw BJ, Cyster JG. Transcriptional regulation of memory B cell differentiation. Nat Rev Immunol. Nature Publishing Group; 2021;21:209–20.

44. Zhang Y, Garcia-Ibanez L, Ulbricht C, Lok LSC, Pike JA, Mueller-Winkler J, et al. Recycling of memory B cells between germinal center and lymph node subcapsular sinus supports affinity maturation to antigenic drift. Nat Commun. Nature Publishing Group; 2022;13:2460.

45. Mesin L, Schiepers A, Ersching J, Barbulescu A, Cavazzoni CB, Angelini A, et al. Restricted Clonality and Limited Germinal Center Reentry Characterize Memory B Cell Reactivation by Boosting. Cell. 2020;180:92-106.e11.

46. Reusch L, Angeletti D. Memory B-cell diversity: From early generation to tissue residency and reactivation. European Journal of Immunology. 2023;53:2250085.

47. Palm A-KE, Henry C. Remembrance of Things Past: Long-Term B Cell Memory After Infection and Vaccination. Frontiers in Immunology [Internet]. 2019 [cited 2023 Sep 29];10. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2019.01787

48. Manara F, Jay A, Odongo GA, Mure F, Maroui MA, Diederichs A, et al. Epigenetic Alteration of the Cancer-Related Gene TGFBI in B Cells Infected with Epstein–Barr Virus and Exposed to Aflatoxin B1: Potential Role in Burkitt Lymphoma Development. Cancers. Multidisciplinary Digital Publishing Institute; 2022;14:1284.

49. Hernandez-Vargas H, Gruffat H, Cros MP, Diederichs A, Sirand C, Vargas-Ayala RC, et al. Viral driven epigenetic events alter the expression of cancer-related genes in Epstein-Barr-virus naturally infected Burkitt lymphoma cell lines. Sci Rep. 2017;7:5852.

50. Zhang T, Ma J, Nie K, Yan J, Liu Y, Bacchi CE, et al. Hypermethylation of the tumor suppressor gene PRDM1/Blimp-1 supports a pathogenetic role in EBV-positive Burkitt lymphoma. Blood Cancer Journal. Nature Publishing Group; 2014;4:e261–e261.

51. Stanland LJ, Luftig MA. The Role of EBV-Induced Hypermethylation in Gastric Cancer Tumorigenesis. Viruses. 2020;12:1222.

52. Sinclair AJ. Could Changing the DNA Methylation Landscape Promote the Destruction of Epstein-Barr Virus-Associated Cancers? Frontiers in Cellular and Infection Microbiology [Internet]. 2021 [cited 2023 Jun 2];11. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2021.695093

53. Ghosh Roy S, Robertson ES, Saha A. Epigenetic Impact on EBV Associated B-Cell Lymphomagenesis. Biomolecules. Multidisciplinary Digital Publishing Institute; 2016;6:46.

54. Saha A, Jha HC, Upadhyay SK, Robertson ES. Epigenetic silencing of tumor suppressor genes during in vitro Epstein–Barr virus infection. Proceedings of the National Academy of Sciences. Proceedings of the National Academy of Sciences; 2015;112:E5199–207.

55. Asmar F, Punj V, Christensen J, Pedersen MT, Pedersen A, Nielsen AB, et al. Genome-wide profiling identifies a DNA methylation signature that associates with TET2 mutations in diffuse large B-cell lymphoma. Haematologica. 2013;98:1912–20.

56. Rosikiewicz W, Chen X, Dominguez PM, Ghamlouch H, Aoufouchi S, Bernard OA, et al. TET2 deficiency reprograms the germinal center B cell epigenome and silences genes linked to lymphomagenesis. Science Advances. American Association for the Advancement of Science; 2020;6:eaay5872.

57. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proc Natl Acad Sci U S A. 2003;100:9991–6.

58. Scott DW, Wright GW, Williams PM, Lih C-J, Walsh W, Jaffe ES, et al. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. Blood. 2014;123:1214–7.

59. Yan W-H, Jiang X-N, Wang W-G, Sun Y-F, Wo Y-X, Luo Z-Z, et al. Cell-of-Origin Subtyping of Diffuse Large B-Cell Lymphoma by Using a qPCR-based Gene Expression Assay on Formalin-Fixed Paraffin-Embedded Tissues. Front Oncol. 2020;10:803.

60. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.

61. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, et al. De novo identification of differentially methylated regions in the human genome. Epigenetics & Chromatin. 2015;8:6.

62. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. Nature Publishing Group; 2017;14:407–10.

63. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016;32:1446–53.

64. Abante J, Kambhampati S, Feinberg AP, Goutsias J. Estimating DNA methylation potential energy landscapes from nanopore sequencing data. Sci Rep. Nature Publishing Group; 2021;11:21619.

65. Gu Z, Hübschmann D. rGREAT: an R/bioconductor package for functional enrichment on genomic regions. Bioinformatics. 2023;39:btac745.

Figures

Figure 1. BL genomes contain global demethylation and localized hypermethylation.

A) Global averaged methylation of normal centroblast samples (n=6) and EBV-positive BL (n=130). Methylation values from individual EPIC array probes were binned into 1Mbp windows and the average of each bin is shown on a heat scale. Data are arranged in concentric rings starting from the outermost ring: normal centroblasts, the difference in methylation levels between EBV-positive BL and centroblasts, EBV-positive BL. **B**) Global averaged methylation of normal centroblast samples (n=6) and EBV-negative BL (n=88). Methylation values from individual EPIC array probes were binned into 1Mbp windows and the average of each bin is shown on a heat scale. Data are arranged in concentric rings starting from the outermost ring: normal centroblasts, the difference between EBV-negative BL and normal centroblasts, EBV-negative BL **C)** A comparison of the number of significant DMPs (|abs(logFC)|> 0.25 and Q < 0.05) from Illumina EPIC array that were hypermethylated and hypomethylated among each of EBV-positive BL(n=130) and EBV-negative BL(n=88) when compared to normal centroblast (n=6) samples. **D)** The chromatin state associated with the DMPs in **(C)** shown as a total number of probes with differential methylation status (DMPs) in BL versus normal centoblasts. **E)** The averaged methylation of samples across DMRs shared between EBV-positive (n=130) and EBV-negative (n=88) samples when compared to normal centroblast (n=6) samples.

Figure 2. Methylation patterns in BL are associated with EBV status

A) The chromatin state associated with significant DMPs from the EPIC array that were hypermethylated and hypomethylated from comparing EBV-positive BL (n=130) against EBV-negative BL (n=88) shown as a number of all significant DMPs. The number of EPIC array probes contained within each chromatin region that were differentially methylated between BL and CB (from Figure 1C) is displayed as the background reference. **B)** The averaged methylation of samples across DMRs identified from the

EPIC array comparing EBV-positive (n=130) and EBV-negative BL (n=88). Normal centroblast samples (n=6) are shown as a reference. **C)** The heatmap on the left represents averaged methylation of samples across DMRs identified from the EPIC array comparing EBV-positive (n=130) and EBV-negative (n=88) BL that were significantly ($R^2 > 0.4$ and $Q < 0.01$) correlated with gene expression as quantified by RNA-Seq analysis. The heatmap on the right indicates the expression of the genes associated with each DMR. Rows in both heatmaps are in the same order with each row depicting a DMR and the expression of its associated gene.

Figure 3. Identification of distinct epitypes in BL.

A) Heatmap depicting the optimal 60 probes from the EPIC array which achieved the greatest accuracy for classifying samples into one of the two epitypes (linear predictor score, LPS class). Samples are ordered based on the probability score associated with belonging to the HyperBL epitype. Alluvial plots showing the distribution of epitype membership by EBV status **(B),** age group **(C),** and genetic subgroup **(D).** All 218 samples with the available EPIC array profiling are included in the analysis.

Figure 4. BL epitypes have distinct methylation patterns.

A) Global averaged methylation of normal centroblast samples (n=6) and HyperBL (n=102). Methylation values from individual EPIC array probes were binned into 1Mbp windows and the average of each bin is shown on a heat scale. Data are arranged in concentric rings starting from the outermost ring: normal centroblast samples, the difference in methylation levels between HyperBL and centroblasts, HyperBL **B)** Global averaged methylation of normal centroblast samples (n=6) and HypoBL (n=88). Methylation values from individual EPIC array probes were binned into 1Mbp windows and the average of each bin is shown on a heat scale. Data are arranged in concentric rings starting from the outermost ring: normal centroblast samples, the difference in methylation levels between HypoBL and centroblasts, HypoBL **C)** The chromatin state associated with significant DMPs from the EPIC array that were hypermethylated and hypomethylated from comparing HyperBL(n=102) against HypoBL(n=88) shown as a

number of all significant DMPs (|abs(logFC)|> 0.25 and Q < 0.05). The number of EPIC array probes contained within each chromatin region that were differentially methylated between BL and CB (from Figure 1C) is displayed as the background reference. **D)** The averaged methylation of samples across DMRs identified from the EPIC array comparing HyperBL(n=102) and HypoBL (n=88). Normal centroblast samples (n=6) are shown as a reference and unclassified samples (n=28) are shown for comparison. **E)** The heatmap on the left represents averaged methylation of samples across DMRs identified from the EPIC array comparing HyperBL (n=102) and HypoBL (n=88) that were significantly correlated with gene expression. The heatmap on the right indicates the expression of the genes associated with each DMR. Rows in both heatmaps are in the same order with each row depicting a DMR and the expression of its associated gene. Unclassified samples (n=28) are included for comparison **F)** Dot plot depicting the top 15 enriched families of TFs with binding sites identified within the DMRs based on epitype. The x axis depicts the different families of TFs while the y axis depicts the -log10(p-value) of each TF.

Figure 5. Mutational profiles associated with the BL epitypes.

A) Oncoplot of coding mutations identified by whole-genome sequencing in the genes determined to be associated with BL(29) and their occurrence across epitypes. The percentages on the left indicate frequency of coding mutations of each specific gene across all samples. Each column of the oncoplot represents an individual sample. The mutations are colored according to their type. Grey tiles on the oncoplot represent absence of mutations of the specific gene. Samples (n=190) within each epitype are ordered based on EBV status to highlight key differences. **B)** Box-whisker plots showing the mutation burden across HyperBL (n=102), HypoBL (n=88), and unclassified (n=28) samples. In order from left to right the plots show the total number of coding mutations, driver mutations, and genome-wide mutation load as identified from the whole-genome sequencing. Samples were subject to a Wilcoxon rank sum test (*P < 0.05, **P < 0.01, ***P<0.001, P values at or above 0.05 are not shown) **C)** Estimated number of mutations per COSMIC signatures SBS1, SBS5, and SBS9 in BL tumors (n=218)

stratified by epitype. Each point represents individual sample, and the y-axis shows total number of mutations associated with that signature. Samples were subject to a Wilcoxon rank sum test (*P < 0.05, **P < 0.01, ***P<0.001, P values at or above 0.05 are not shown).

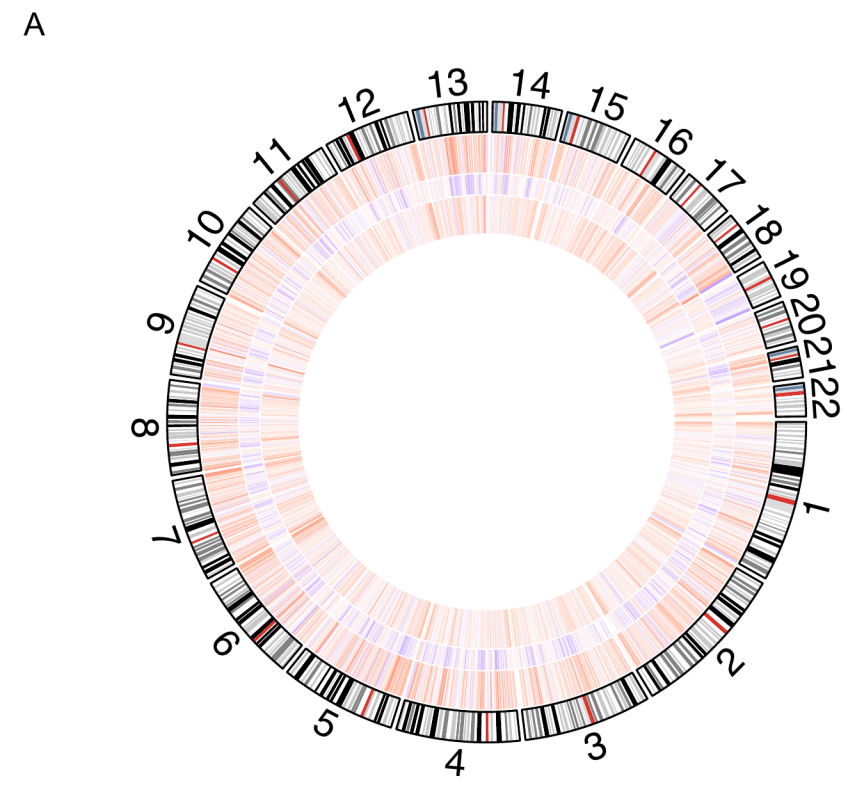
Figure 6. Patterns of gene expression associated with epitypes membership.

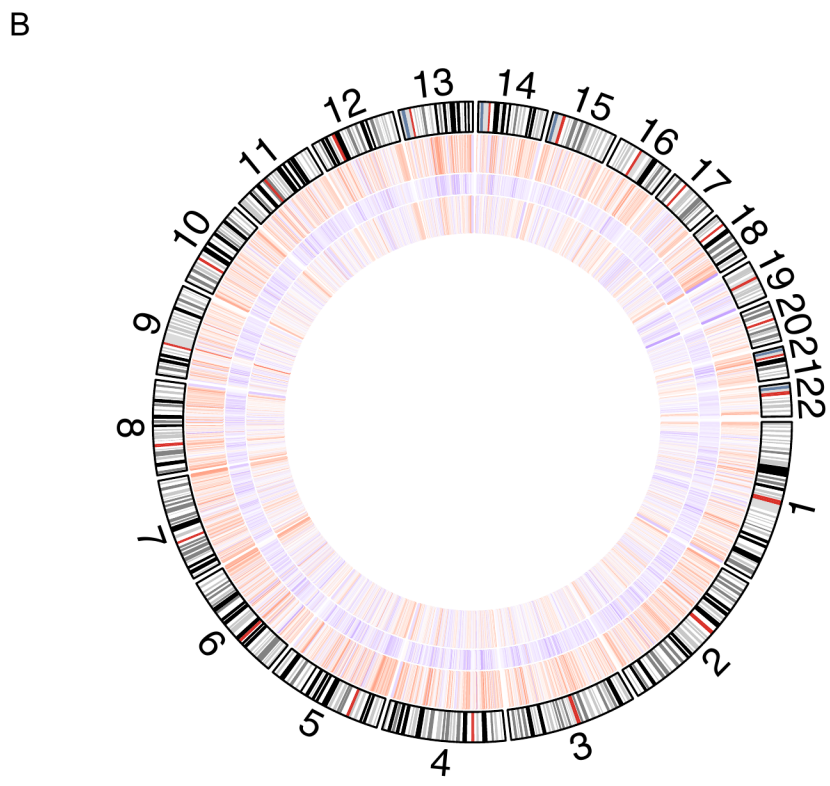
A) The heatmap displays the 213 differentially expressed genes between epitypes, with rows representing differentially expressed genes and columns representing samples. Rows and columns are clustered using Manhattan distances. The epitype membership and EBV status of each sample is shown at the top. Samples are split based on epitype with HyperBL (n=102) cases on the left, HypoBL (n=88) cases in the middle and unclassified (n=28) cases on the right. **B)** Heatmap representing the expression of 32 gene sets from signatureDB with significant (Q <= 0.0025) differences between epitypes. Samples from 218 patients were clustered and ordered on their expression of genes within each gene set. Rows represent the gene sets and columns represent samples. Row and columns were clustered using Manhattan distances.

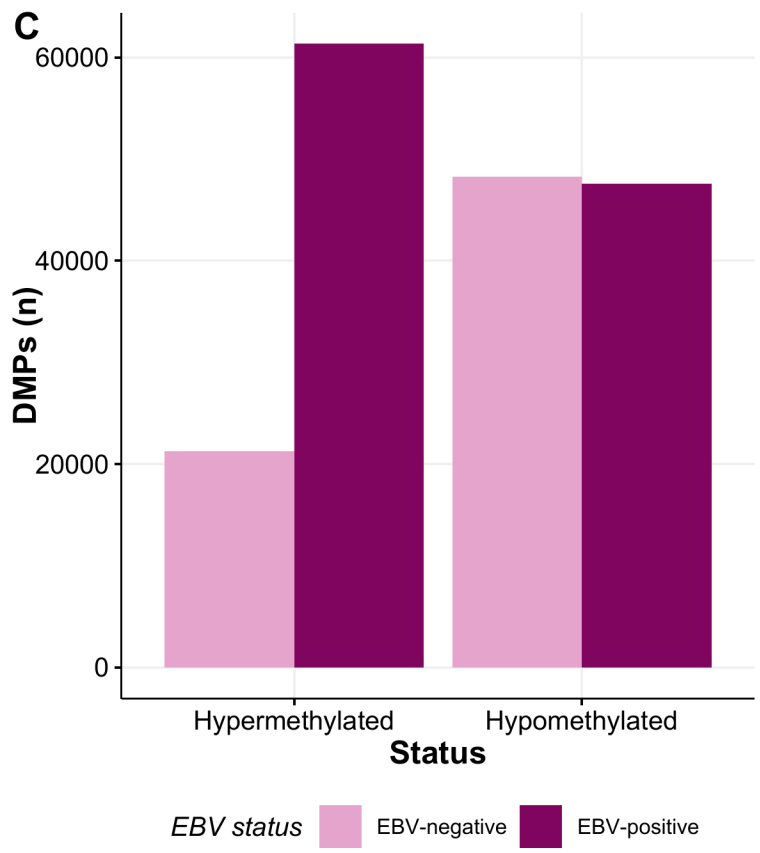
Figure 7. Survival outcomes of adult BL patients stratified by epitype.

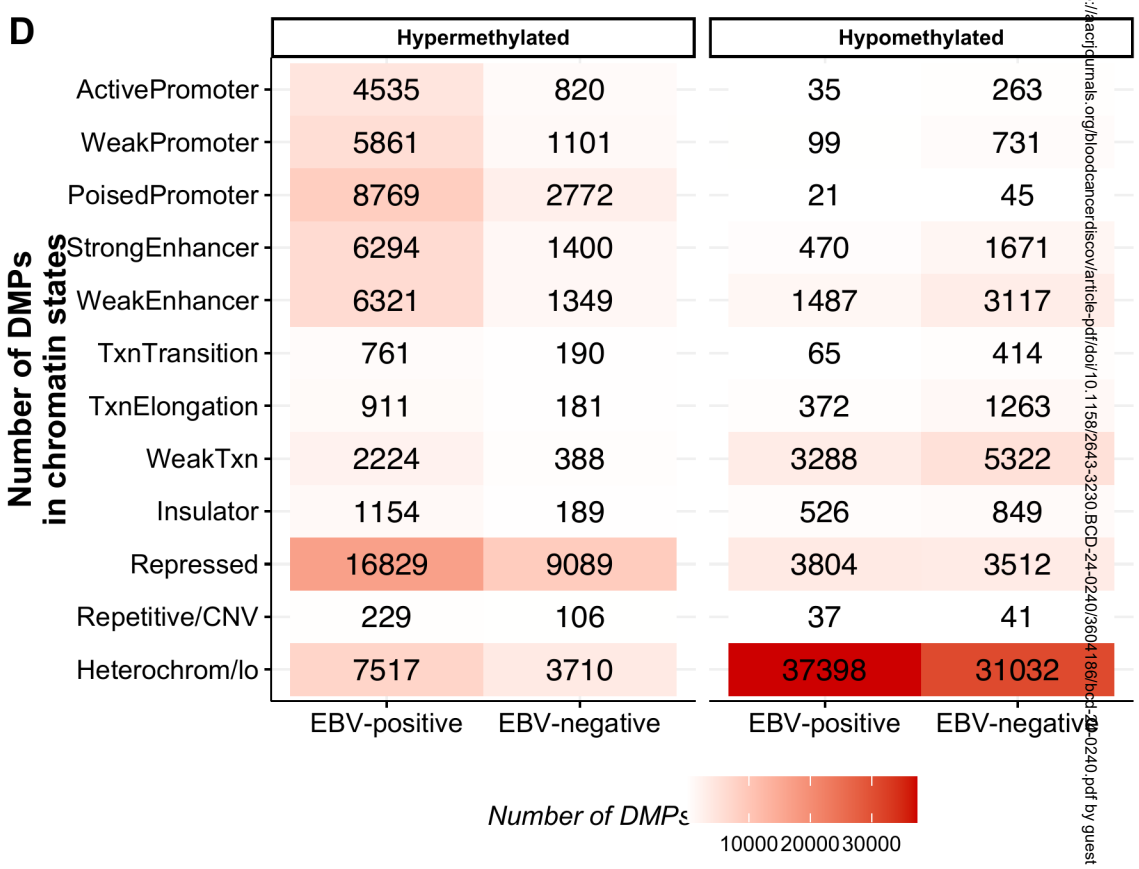
Outcomes of adult BL when stratified by epitype. Kaplan-Meier survival analyses was conducted within the adult cases, for whom the follow-up data was most complete. HyperBL patients were compared to HypoBL patients to analyze progression-free **(A)** and overall **(B)** survival, including the unclassified cases. All times are shown in years. Log rank p-value is shown, and the pairwise comparisons were conducted separately across the indicated groups. The risk tables show the number of patients at the specified timepoint.

Figure 1









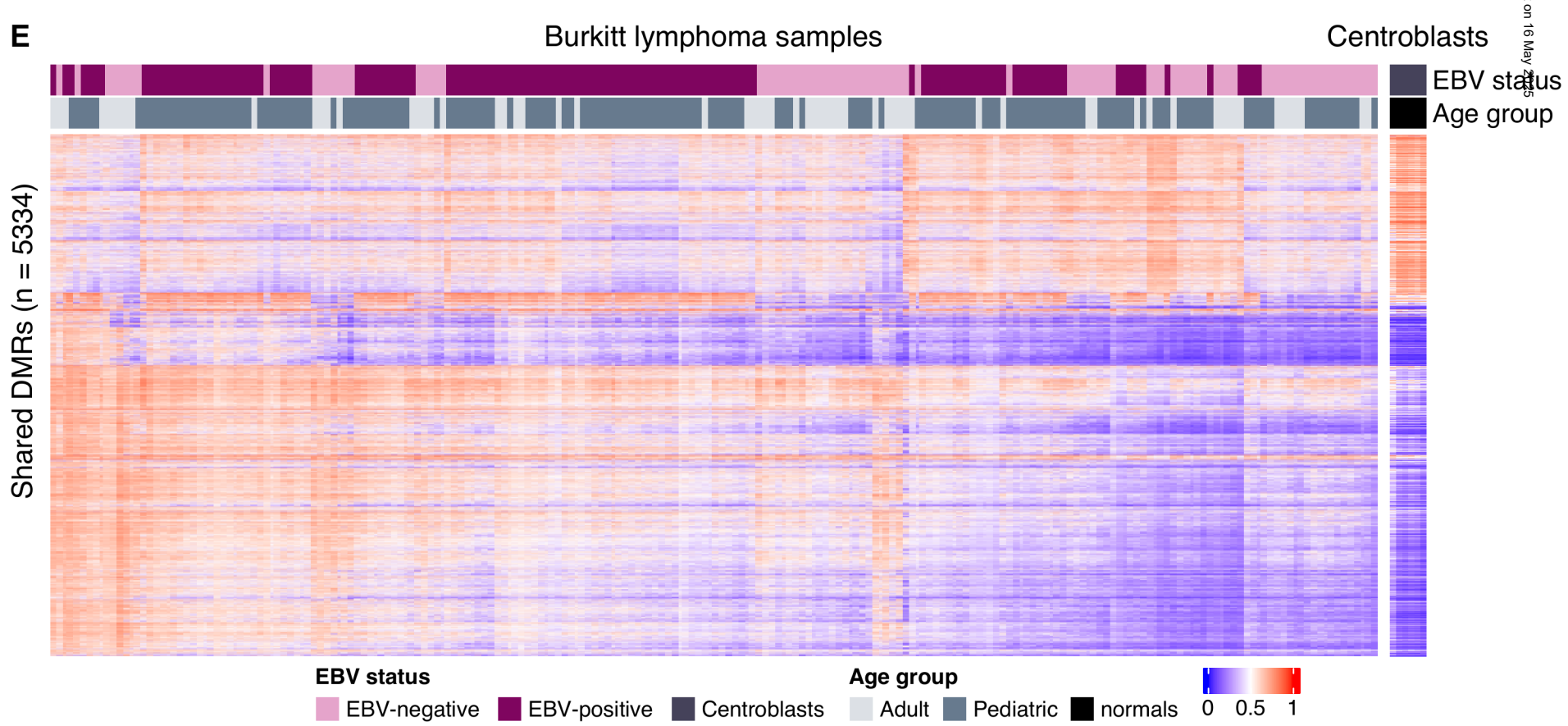
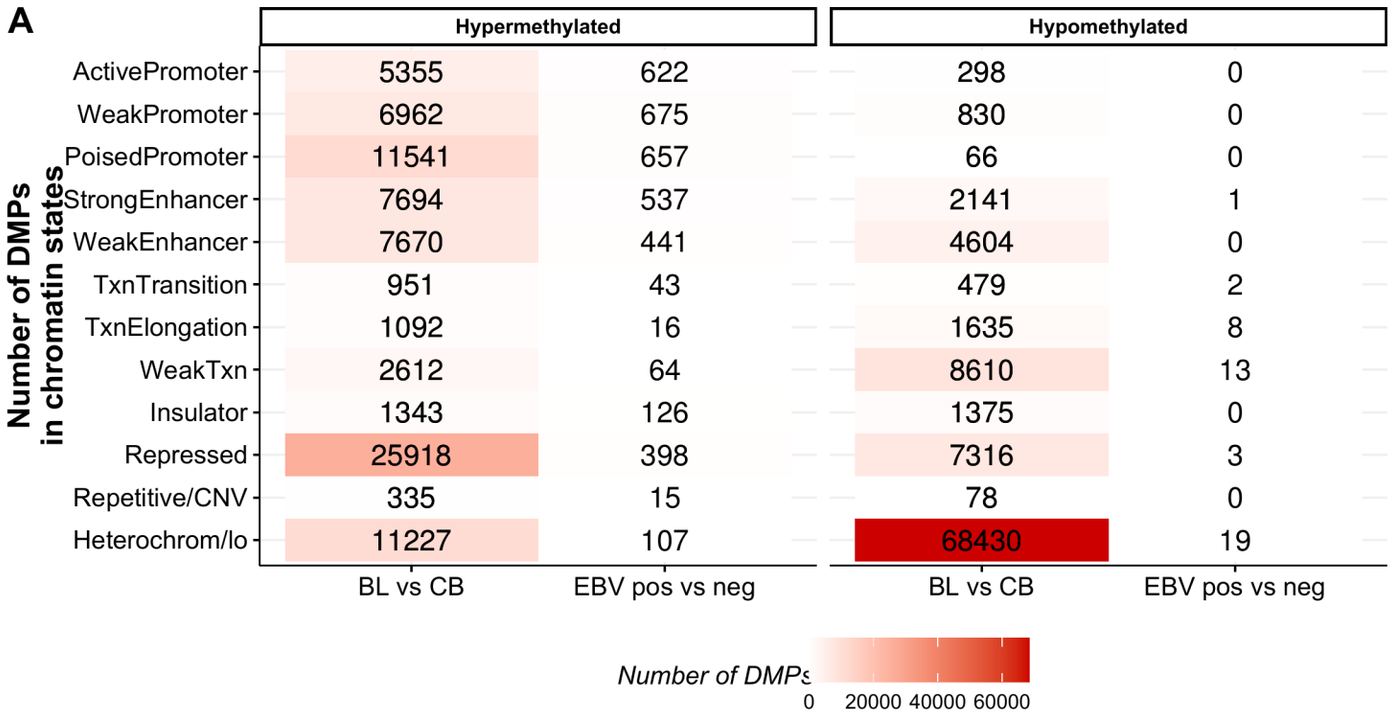
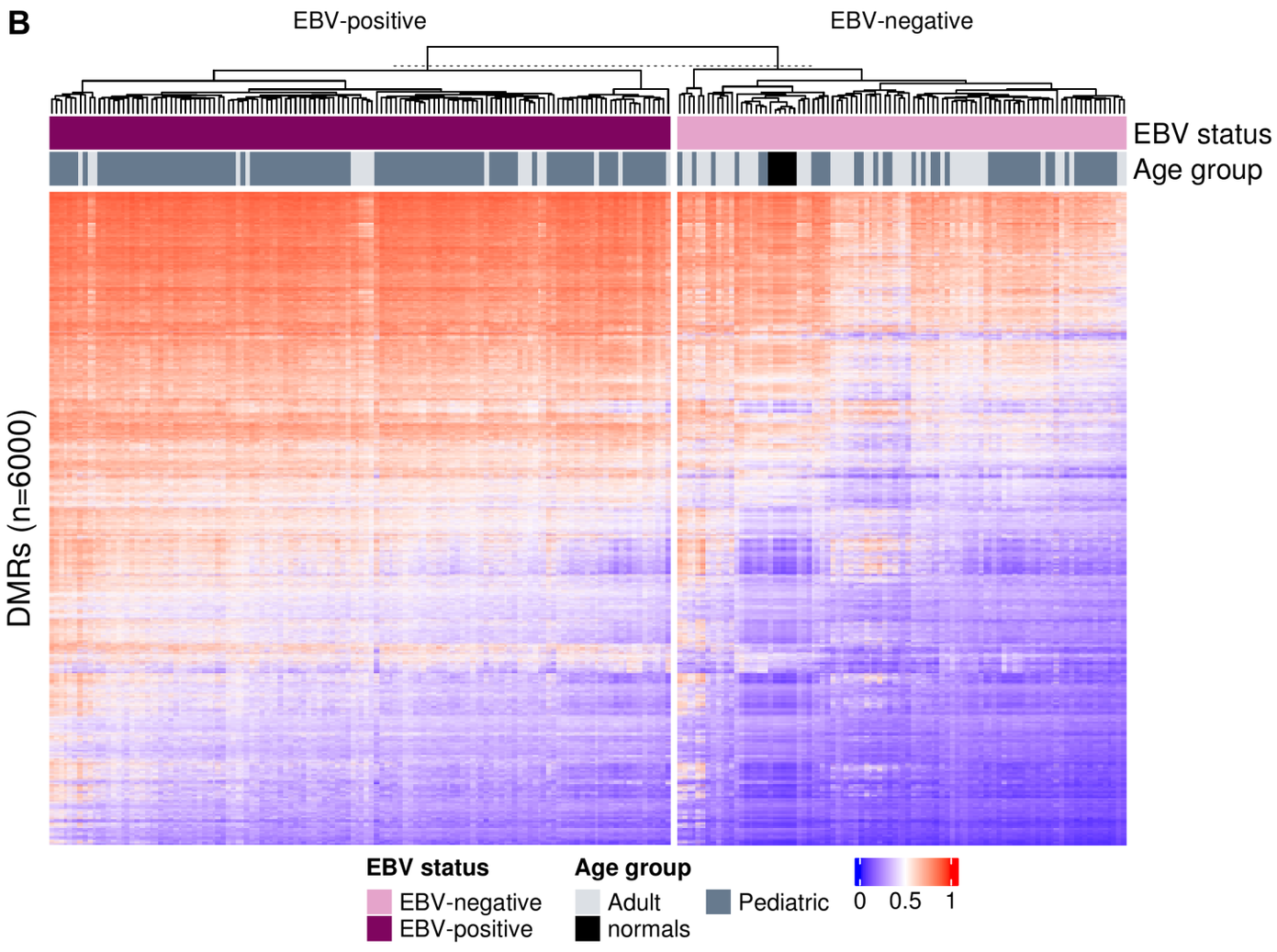


Figure 2





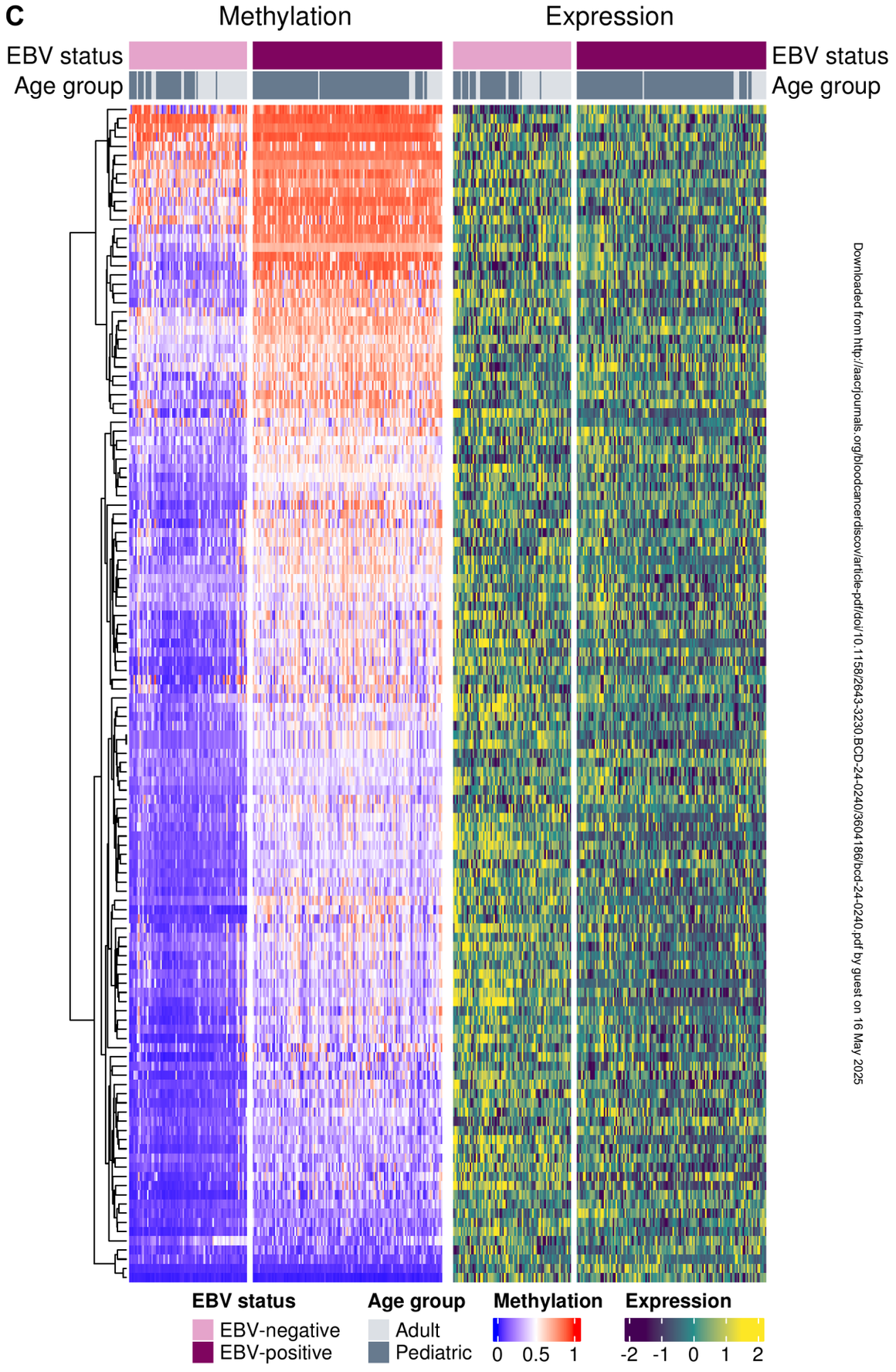


Figure 3

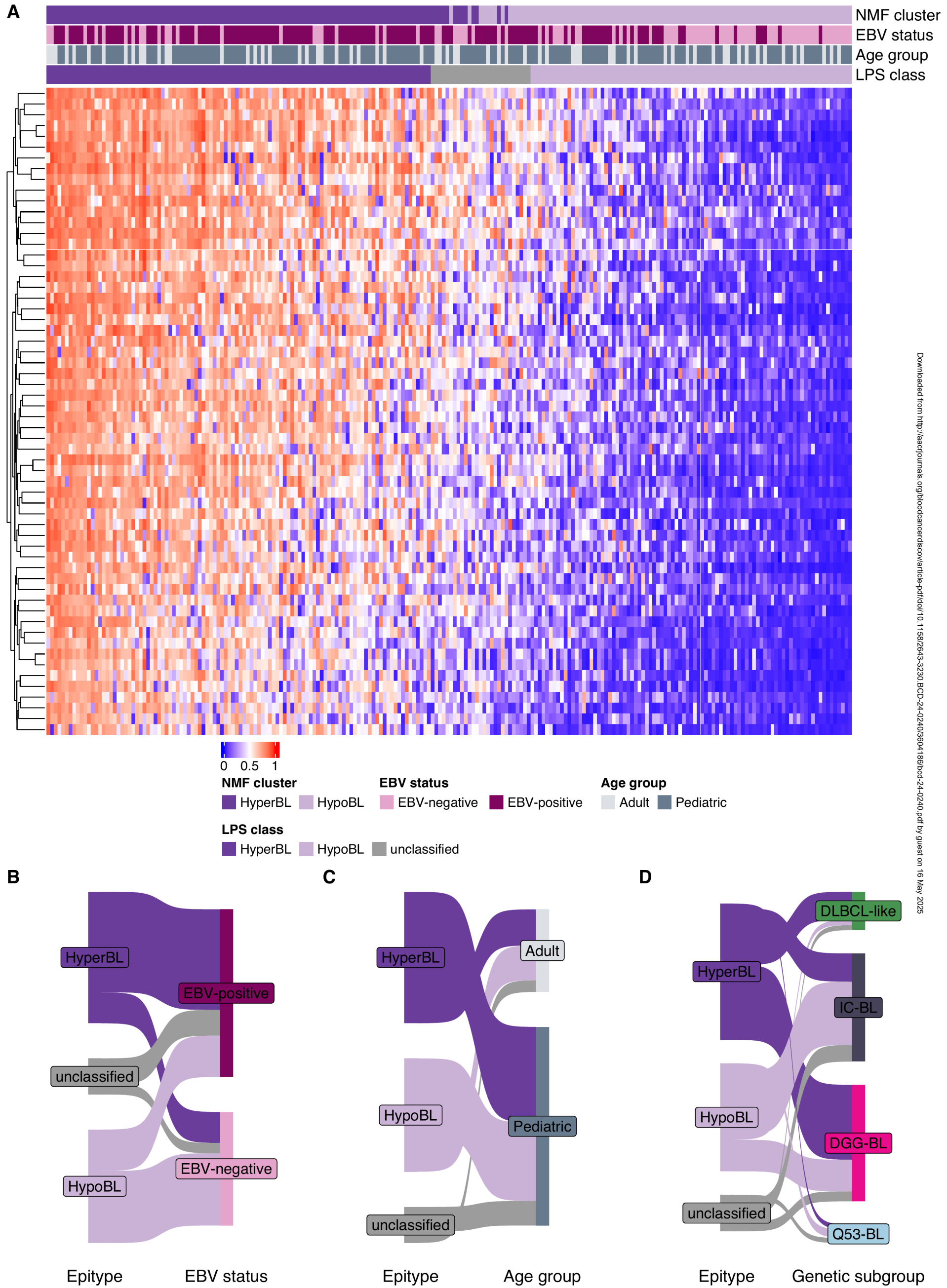


Figure 4

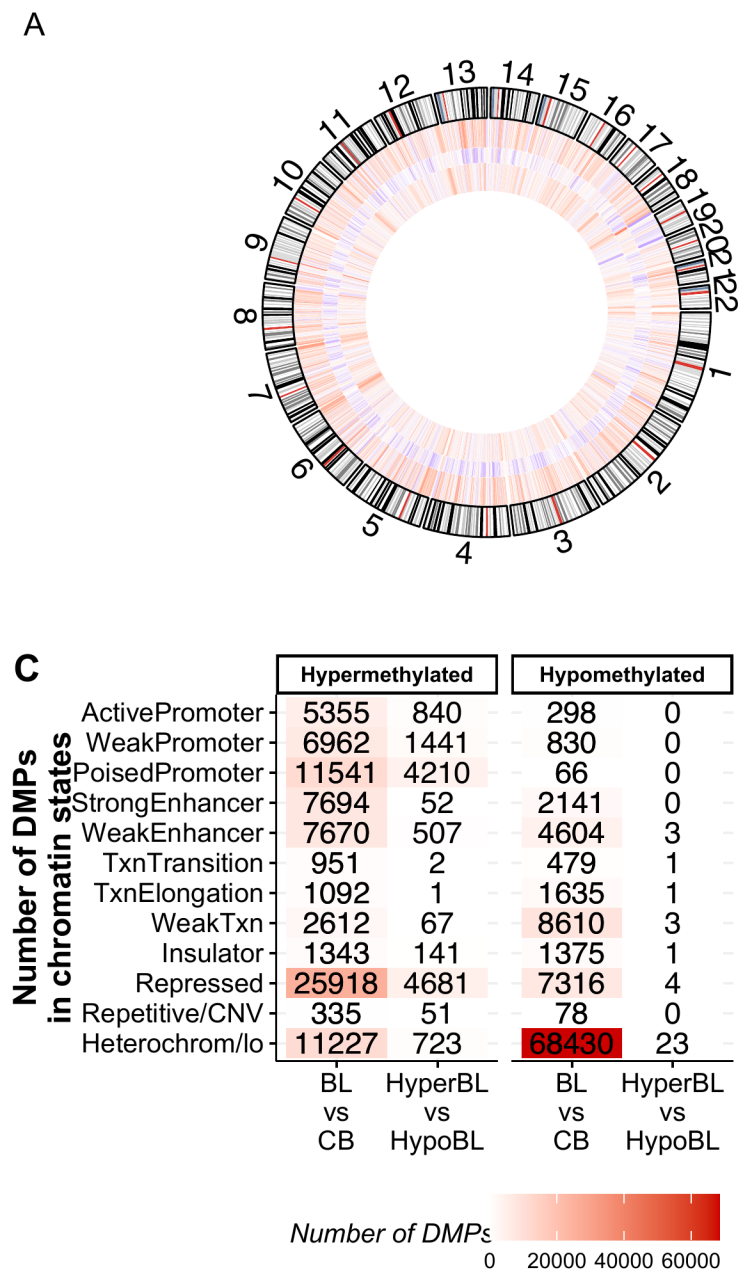
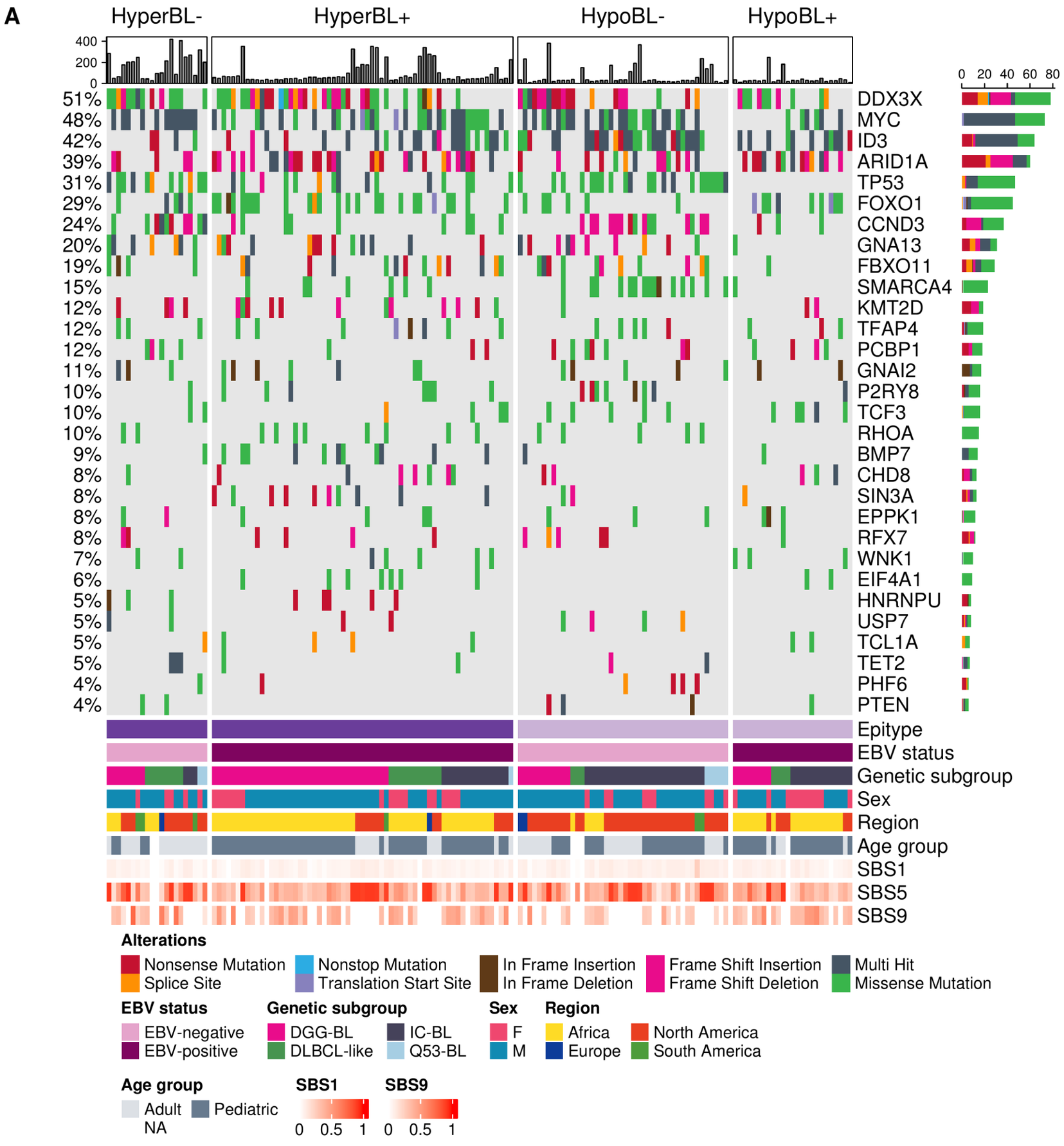


Figure 5



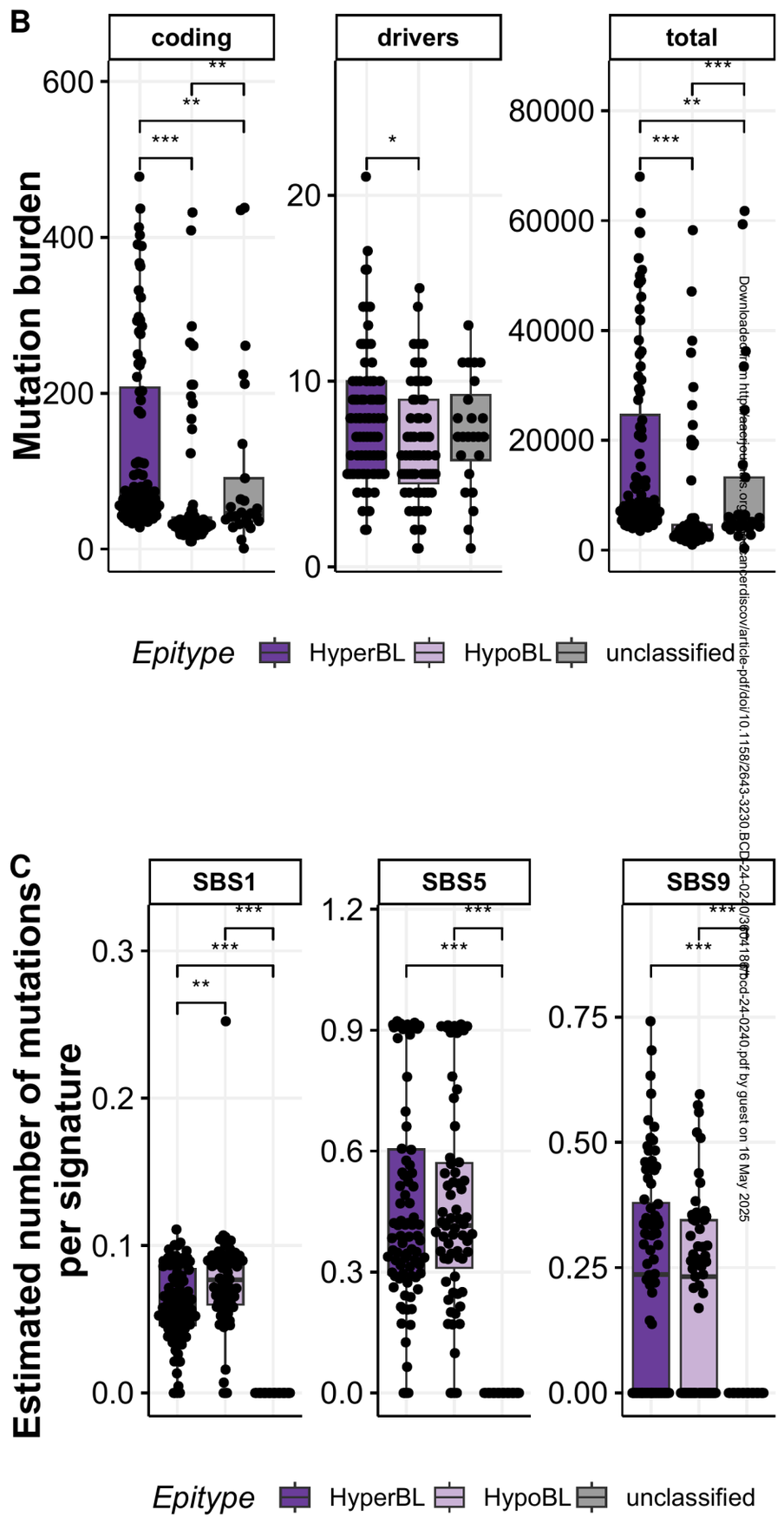
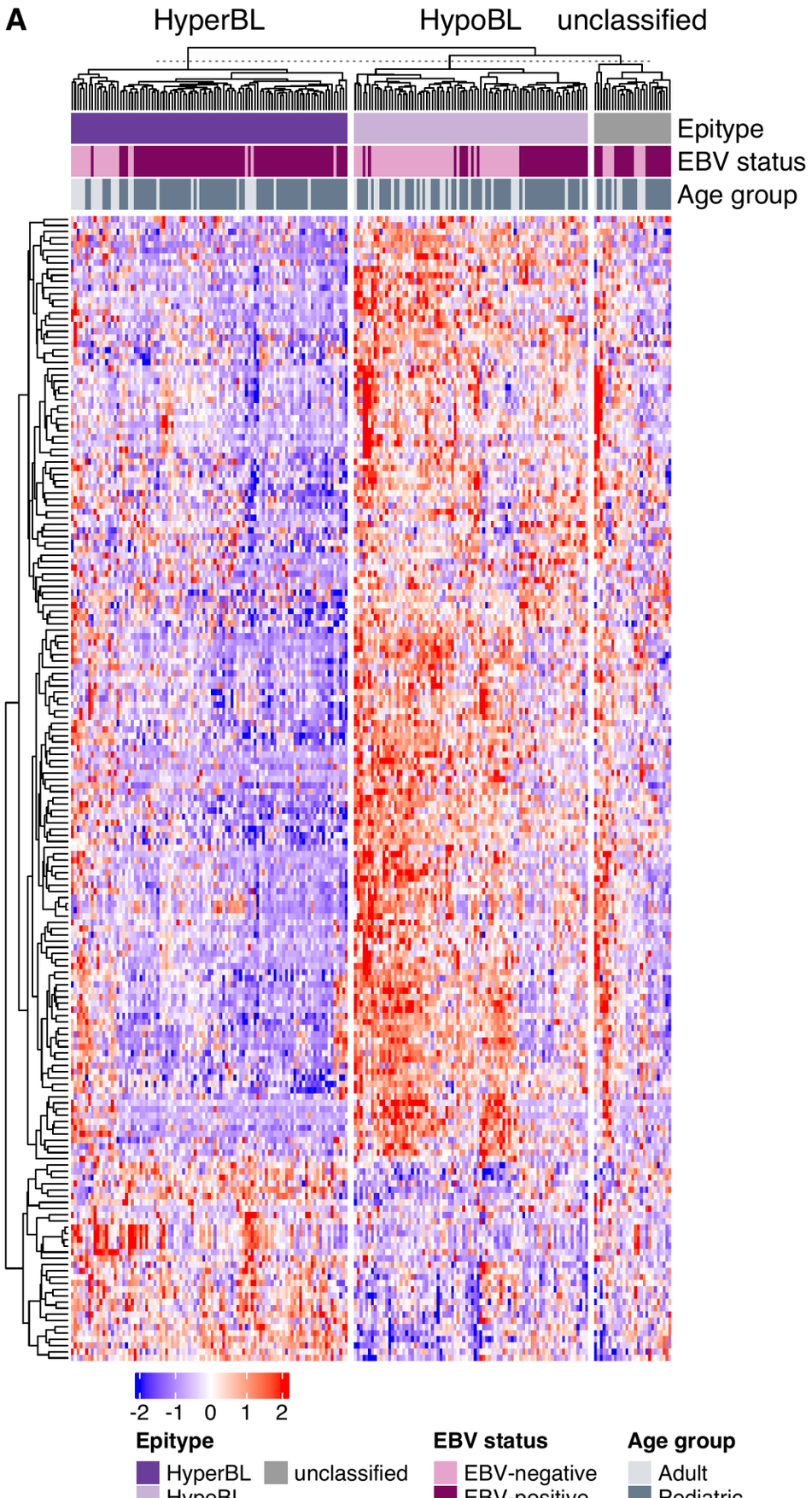


Figure 6



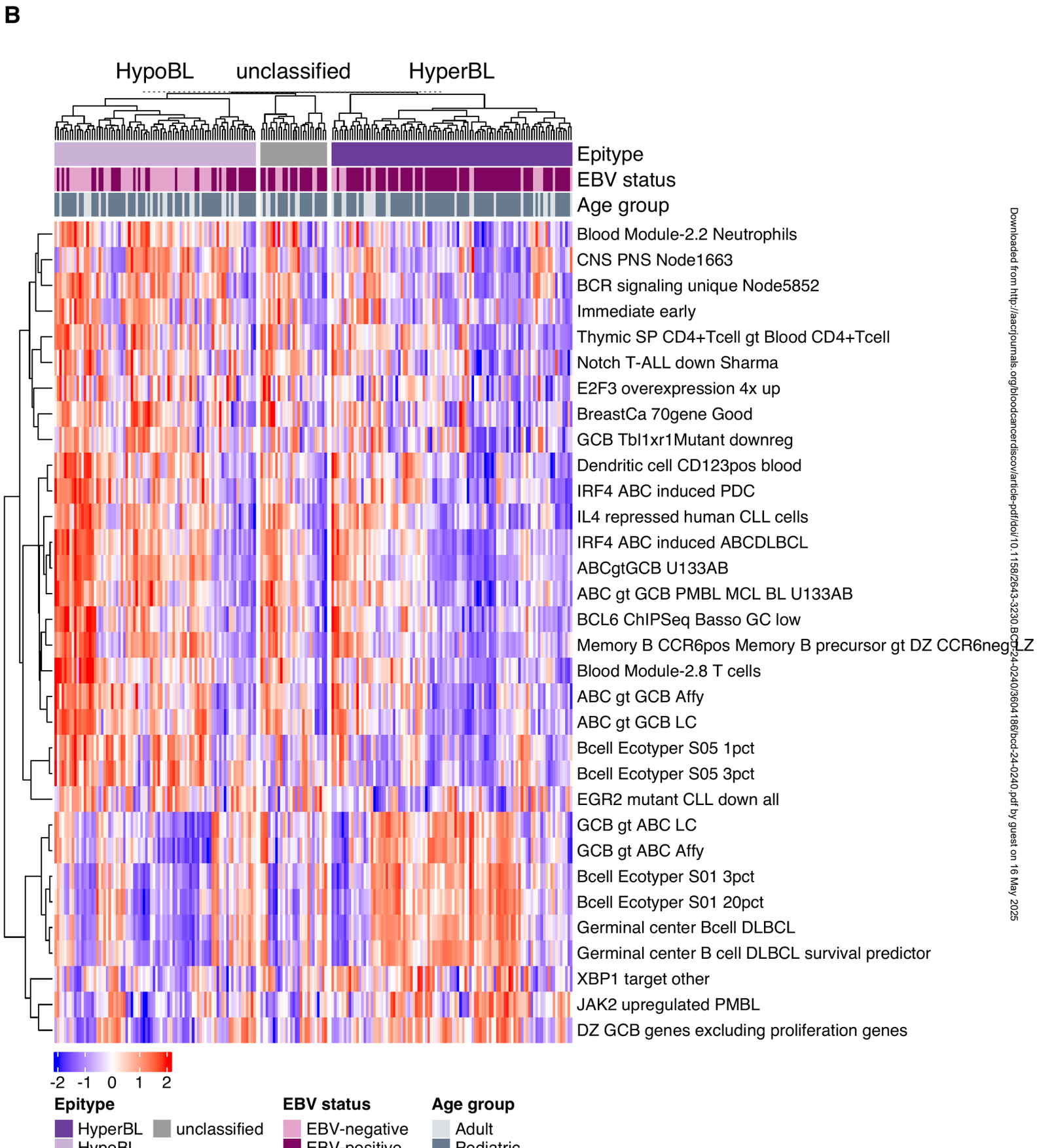


Figure 7