Epstein-Barr virus and immune status imprint the immunogenomics of non-Hodgkin lymphomas occurring in immune-suppressed environments

Marine Baron,[1,2*] Karim Labreche,[3*] Marianne Veyri,[4+] Nathalie Désiré,[3+] Amira Bouzidi,[5] Fatou Seck-Thiam,[3] Frédéric Charlotte,[6] Alice Rousseau,[1] Véronique Morin,[1] Cécilia Nakid-Cordero,[1] Baptiste Abbar,[1] Alberto Picca,[1] Marie Le Cann,[7] Noureddine Balegroune,[2] Nicolas Gauthier,[2] Ioannis Theodorou,[8] Mehdi Touat,[9] Véronique Morel,[2] Franck Bielle,[10] Assia Samri,[1] Agusti Alentorn,[9] Marc Sanson,[9] Damien Roos-Weil,[2] Corinne Haioun,[11] Elsa Poullot,[12] Anne Langlois de Septenville,[13] Frédéric Davi,[13] Amélie Guihot,[1] Pierre-Yves Boelle,[3] Véronique Leblond,[2] Florence Coulet,[14] Jean-Philippe Spano,[4#] Sylvain Choquet,[2#] and Brigitte Autran[1#] on behalf the IDeATIon study group. IDeATIon study group: Baptiste Abbar,[4] Isabelle Brocheriou,[6] Jacques Cadranel,[15] Jérôme Denis,[5] Erell Guillerm,[13] Ahmed Ibdaih,[8] Stéphanie Jouannet,[5] Jean-Marc Lacorte,[5] Anne-Geneviève Marcelin,[14] Alberto Picca,[2] Kahina Belkhir[4] and Cécilia Nakid-Cordero[2]

[1]Sorbonne Université, INSERM U1135, Center for Immunology and Infectious Diseases (CIMI), Department of Immunology, AP-HP, Hôpital Pitié-Salpêtrière, Paris; [2]Sorbonne Université, Department of Clinical Hematology, AP-HP, Hôpital Pitié-Salpêtrière, Paris; [3]Sorbonne Université, CinBioS, UMS 37 PASS Production de données en Sciences de la vie et de la Santé, INSERM, Paris; [4]Sorbonne Université, INSERM, Pierre et Louis Institute of Epidemiology and Public Health, Theravir Team, Department of Medical Oncology, AP-HP, Hôpital Pitié-Salpêtrière, Paris; [5]Sorbonne Université, INSERM, Research Unit on Cardiovascular and Metabolic Disease UMR ICAN, Department of Endocrine Biochemistry and Oncology, AP-HP, Hôpital Pitié-Salpêtrière, Paris; [6]Sorbonne Université, Department of Anatomy and Pathologic Cytology, AP-HP, Hôpital Pitié-Salpêtrière, Paris; [7]Department of Clinical Hematology, AP-HP, Hôpital Kremlin Bicêtre, Le Kremlin; [8]Department of Immunology, Hôpital Robert Debré, Paris; [9]Sorbonne Université, INSERM, CNRS, Brain and Spine Institute, ICM, Department of Neurology 2-Mazarin, AP-HP, Hôpital Pitié-Salpêtrière, Paris; [10]Sorbonne Université, Department of Neuropathology, AP-HP, Hôpital Pitié-Salpêtrière, Paris; [11]Lymphoid Malignancies Unit, AP-HP, Mondor Hospital, Créteil; [12]Department of Anatomy and Pathologic Cytology, AP-HP, Mondor Hospital, Créteil; [13]Sorbonne Université, INSERM, Centre de Recherche des Cordeliers, Department of Biological Hematology, AP-HP, Hôpital Pitié-Salpêtrière, Paris; [14]Sorbonne Université, INSERM, Saint-Antoine Research Center, Microsatellites Instability and Cancer, CRSA, Department of Medical Genetics, AP-HP, Pitié-Salpêtrière Hospital, Paris and [15]Sorbonne Université, Department of Pneumology, AP-HP, Hôpital Tenon, Paris, France

[*]MB and KL contributed equally as first authors.

[+]MV and ND contributed equally.

[#]J-PS, SC and BA contributed equally as senior authors.

Correspondence: M. Baron
marine.baron@aphp.fr

Received: September 20, 2023.
Accepted: May 22, 2024.
Early view: June 6, 2024.

https://doi.org/10.3324/haematol.2023.284332

SUPPLEMENTARY METHODS AND MATERIALS

Methods:

Patient cohort

Patients were enrolled from the Hematology or Neuro-Oncology departments at Pitié-Salpêtrière Hospital, Paris, France, or the lymphoid malignancies unit at Henri Mondor Hospital, Créteil, France. The lymphoma subtype was defined on the diagnostic tumor biopsy according to the 2016 WHO classification[1]. The EBV status was assessed using EBV-encoded small RNA (EBER) *in situ* hybridization analysis. In 8 cases where enough tissue was available, an immune-histochemistry analysis of the EBER, LMP-1 and EBNA-2 EBV antigens expression was performed in order to assess the EBV latency status as follows: latency I: EBER+, LMP-1-, EBNA-2-, latency II: EBER+, LMP-1+, EBNA-2- and latency III: EBER+, LMP1+, EBNA-2+.

Whole exome sequencing

WES was performed on tumor genomic DNA from tumor tissue biopsies and on germline DNA from blood. DNA was isolated from fresh frozen or formalin-fixed paraffin-embedded (FFPE) biopsies and from blood using a QIAamp-DNA Mini-Kit, FFPE Tissue-Kit or blood kit (Qiagen) respectively. Libraries were prepared and hybrid-captured using the SeqCap-EZ MedExome-Enrichment Kit (Roche) with 200 ng DNA input. Sequencing was performed on an Illumina Novaseq system with 150-bp paired-end reads. Raw paired-end fastq files were pre-processed using fastp[2] for adapter trimming and quality filtering. The filtered reads were aligned to the hg38 human reference genome using BWA-MEM[3]. The mean sequencing coverage across targeted bases was 200× and 150× for tumour and germline, respectively (**Supplementary Figure 2.A**). Somatic single nucleotide variants (SNVs) and indels were

called using MuTecT[4] and Strelka2[5], respectively. We excluded potential oxidative damage-induced and low base quality mutations using in-house scripts[6]. For two patients with missing germline DNA, we used Mutect2 (v4.1.4.1) and Panel of Normal (PoN) was generated, made from 38 normal samples DNA (derived from healthy tissue). Variants detected (marked PASS, coverage AD >=10 in both normal and tumor, at least 3 tumor variant reads, VAF >=0.05, gnomAD_AF <= 0.0001) were annotated using VEP[7]. Candidate driver genes and significantly mutated genes were defined using OncodriveCLUST[8].

Detection of copy number alteration (CNV)

We employed FACETS[9] tool to accurately infer CNV and tumour purity estimation for the paired samples (n=66). Moreover, GISTIC2[10] was used for the identification of recurrent copy number alterations. By leveraging data from high-throughput genomic profiling technologies, GISTIC2 aids in the detection of focal and arm-level copy number variations that are statistically significant across a cohort of samples. GISTIC2's allow to discern between true somatic events and random fluctuations in copy number.

CCF inference

PyCloneVI[11] was employed to estimate the cancer cell fractions within each tumor sample. FACETS CNV segments and raw variant allele frequencies were provided as input, along with a specified mutation prevalence threshold of 0.05 to exclude low-frequency variants. The 'pyclone-vi fit' and 'pyclone-vi write-results-file' command was executed with default settings.

We determined if a peptide was clonal for each patient by looking at the mutation's CCF that forms the peptide. A peptide is considered clonal when its CCF is 0.8 or higher.

RNA sequencing

RNA was extracted from fresh frozen biopsies using an RNeasy Micro-Kit (Qiagen). Libraries were prepared from 500 ng RNA. After end-repair, A-tailing, ligation and purification, sequencing was performed on an Illumina Novaseq with 150-bp paired-end reads. Reads were aligned to the human hg38 reference genome after an index was generated using STAR v2.7.2[12] by applying per-sample two-pass mapping. The generated BAM files were pre-processed according to GATK v4.1 RNA-seq best practice (SplitNcigarReads and BQSR), duplicates reads were marked then removed using STAR option '--bamRemoveDuplicatesType UniqueIdentical --bamRemoveDuplicatesMate2basesN 15' and Samtools, respectively. Finally, gene counts were obtained using Htseq[9] (RNAseq library size range: 24.8-154.9 Million reads, see **Supplementary Figure 2.B** and transformed into CPM (Counts Per Million) values.

Gene expression and cell type abundance profiling

T cell function included/ *IFN-γ, TNF-α, CD107a, granzyme* B and *perforin,* positive immune regulation included: *ICOS, ICOSL, CD28, CD40, CD40L, OX0, 0X40L* and *4.1BB,* and negative immune regulation included: *PD-1, PD-L1, PD-L2, LAG-3, TIM-3, Galectin9, TIGIT, CD47, IL-10, TGF-β* and *IDO*. Each sample enrichment scores and gene set pairing was computed separately using Single-sample Gene Set Enrichment Analysis (ssGSEA)[13] (https://www.gsea-msigdb.org/gsea/index.jsp).

Differential gene expression (DGE) and gene ontology (GO) enrichment analysis

CPM values were normalized using the Trimmed Mean of M-values normalization method from edgeR package[14]. To filter out low read counts, a CPM threshold of 0.5 was applied, equivalent to a count of 10 for the library sizes in this dataset. Subsequently, the data were log2 normalized (see **Supplementary Figure 2.C**). DGE analysis was performed with edgeR. DE genes were identified using a p-value cut-off of 5% without fold-change cut-off. A GO analysis

was used to predict putative biological functions based on DGE. The DE genes between CNS vs Systemic, EBV+ vs EBV- and ID vs IC groups were inserted to the goana function in limma-R packages[15] with focus on the biological process ontology.

Since the MHC class-I and class-II locus expression was computed using seq2HLA and measured in RPKM, we opted to calculate the *B2M* expression in RPKM as well.

MHC class-I and class-II restricted neoepitope prediction

Neoepitopes were predicted using the Ideation@SiRIC pipeline combining several software packages. First, we used seq2HLA to determine MHC class-I and class-II types [PMID: 23259685] using default parameters for all 31 RNAseq tumors and 68 normal-WES preprocessed fastq files using fastp. Next, somatic mutation-filtered VCF files were annotated by Variant Effect Predictor (Version 99) with default parameters and additionally with the 'Frameshift' and 'wild-type' plugins (from the pVACtools suite [version 3.0.3 PMID: 31907209]). The annotated non-synonymous mutations were extracted for downstream analysis. For each variant, the transcript expression levels (from RNA sequencing data) were then added using vcf-expression-annotator from VAtools v5.0.1. All parameters were then processed with the pVACseq for neoantigen prediction. For each pVACseq run, epitope prediction was done by the NetMHC [PMID: 18463140] NetMHCpan and NetMHCIIpan [PMID: 32406916] algorithms packed in the pVACseq toolkit; epitope length was set to 8–10 amino-acids long for class-I and 15 for class-II presentation with default parameters for all other settings. Predicted neoepitopes were filtered based on coverage >10X, DNA VAF ≥10%, transcript level expression ≥ 0.5 counts per million (CPM), TSL (Transcript Support Level) =1 and median affinity binding ≤500 nM. When RNA samples were unavailable, selection from the WES data used all of these parameters except the transcript level expression. The number

of neoepitopes per tumor was defined as the overall number of unique filtered mutant peptide sequences per tumor.

Neoepitope priority score: the priority score is one way to prioritize neoepitope candidates proposed by pVACseq: each of the following 4 criteria are assigned a rank-ordered value: B= rank of the mutant IC50 binding affinity/ F= rank of fold change between MT and WT allelles/ M= rank of mutant allele expression/ D= rank of tumor DNA VAF. The score is calculated with the following formula: B + F + (M*2) + (D/2), and then converted to a rank.

T cell receptor (TCR) analysis

From the RNAseq fastq files, the productive TCRβ clonotypes numbers and sequences and their frequencies were generated with MixCR V3.0[16] (https://mixcr.readthedocs.io/). Reads were aligned to reference TCR V, D, J and C genes. Each final clonotype was identified by a unique CDR3 sequence and clonotype count. Further analyses were restricted to samples with numbers of productive TCRs >100. The TCR repertoire abundance was defined as the number of unique productive clonotypes. Shannons' entropy index, computed using vdjtools V1.2.1 software[17], was used to compute the TCR repertoire entropy based on the frequency of particular sequences and normalized after division by clonotype log numbers (lower values indicate greater diversity). The clonotypes with frequencies ≥10% were considered to be suggestive of a tumor-neoantigen selection and investigated for known antigen specificities in public dataset (VDJdb)[18].

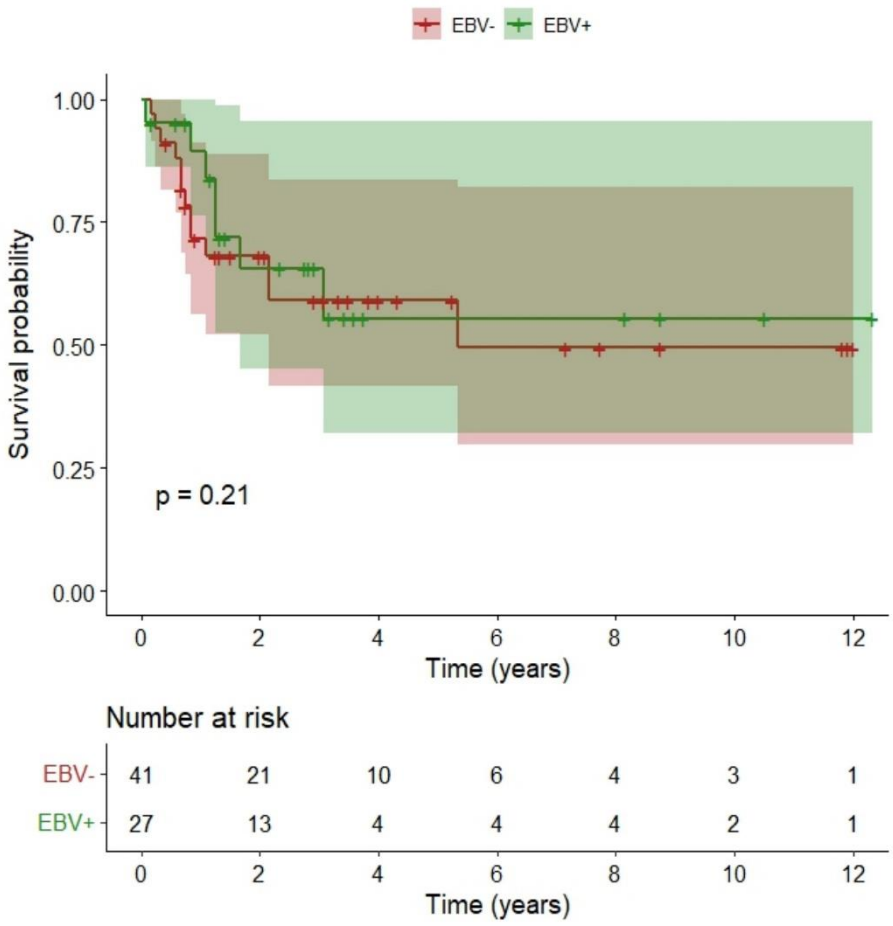
Statistical Analysis

Statistical analysis was performed using R (version 4.2.1). The Fisher exact test was used to compare categorical data and the Wilcoxon test was used to compare quantitative variables between groups. All tests were 2-sided at the threshold of p=0.05. In the case of multiple testing, we controlled FDR at the 5% level. Survival plots were generated using the Kaplan–Meier

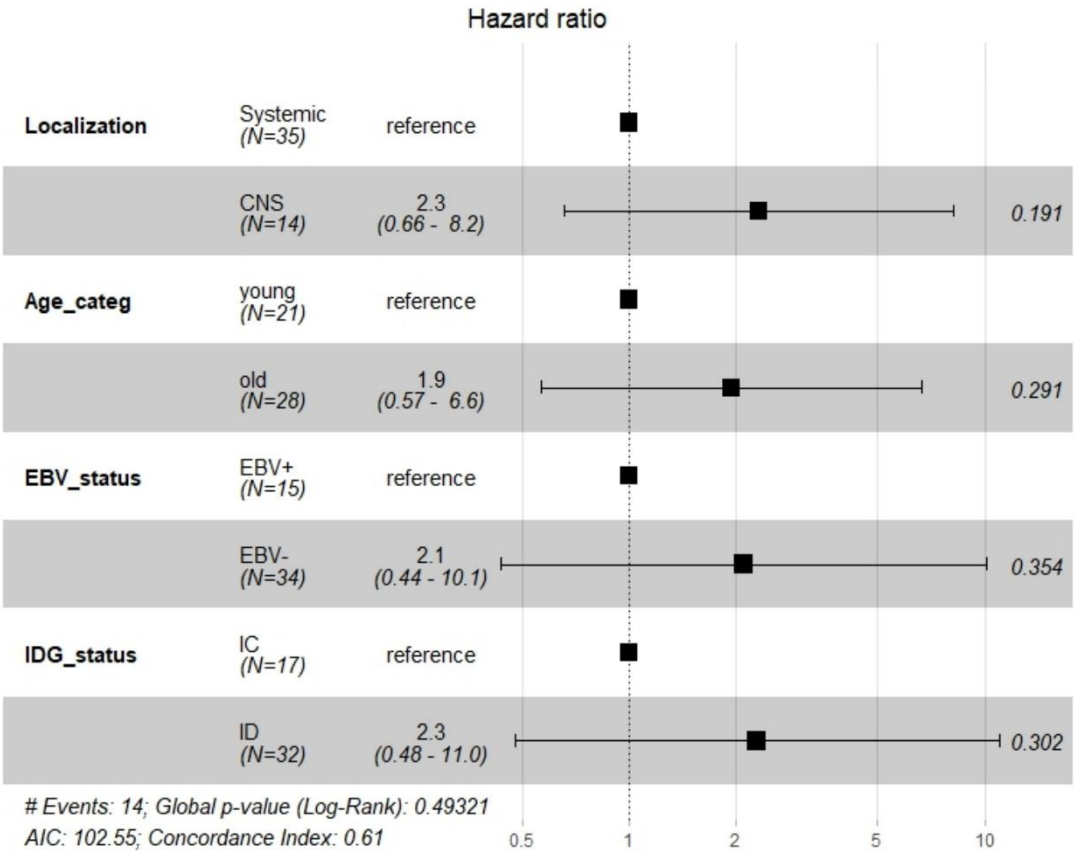
method. The Log-rank test was used to compare survival curves between groups. The Cox proportional-hazards regression model was used to compute hazard ratios summarizing the association between survival and age, EBV status, immune status, and localization. We made use of the “surv_cutpoin” function from survminer R package to determine the optimal TMB cutpoint that corresponds to the most significant relation with survival.

Figures

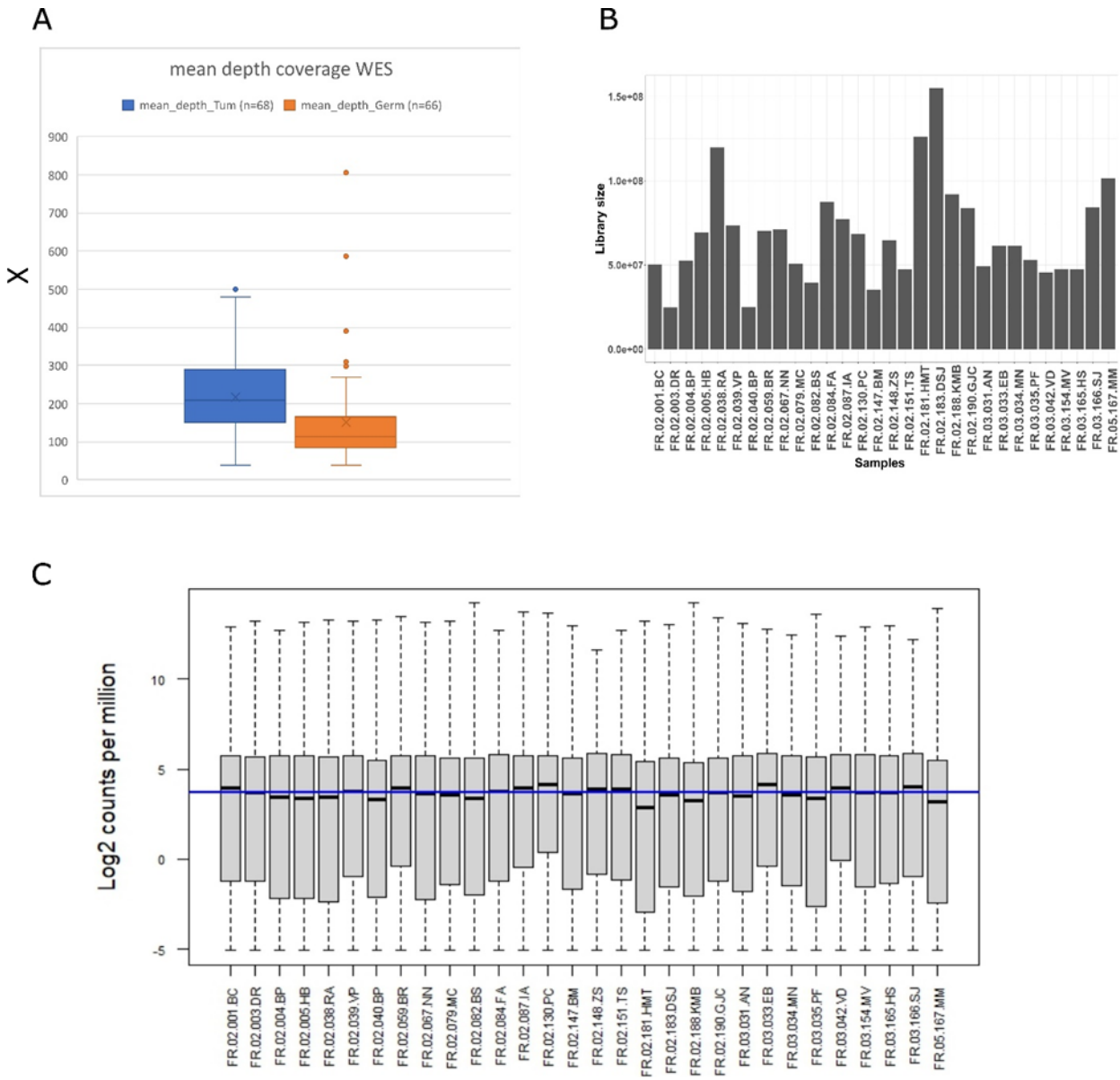
A


EBV-
EBV+
1.00
0.75
0.50
0.25
0.00
Survival probability
p = 0.21
0
2
4
6
8
10
12
Time (years)
Number at risk
EBV- 41 21 10 6 4 3 1
EBV+ 27 13 4 4 4 2 1
0
2
4
6
8
10
12
Time (years)

B

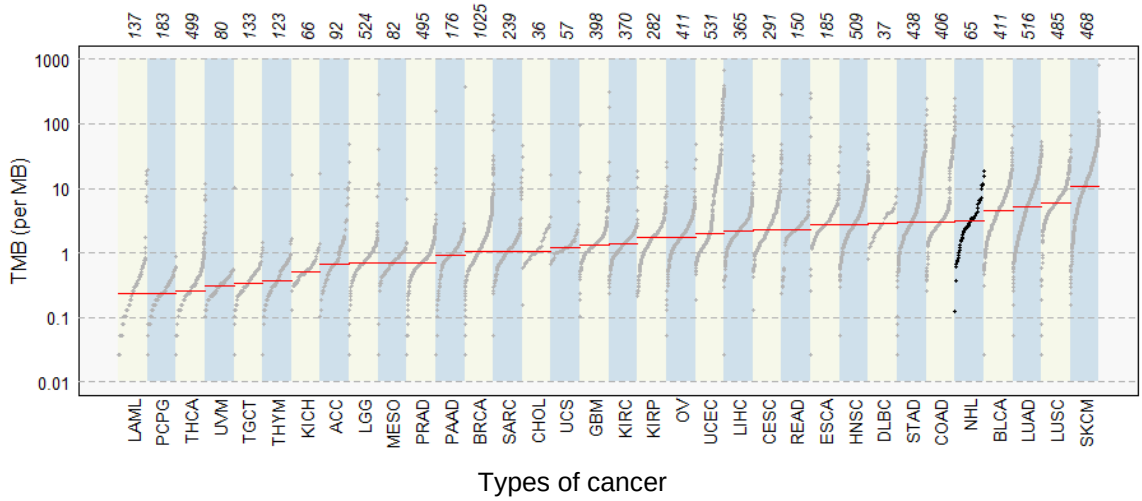

Hazard ratio
Localization
Systemic (N=35)
reference
CNS (N=14)
2.3 (0.66 - 8.2)
0.191
Age_categ
young (N=21)
reference
old (N=28)
1.9 (0.57 - 6.6)
0.291
EBV_status
EBV+ (N=15)
reference
EBV- (N=34)
2.1 (0.44 - 10.1)
0.354
IDG_status
IC (N=17)
reference
ID (N=32)
2.3 (0.48 - 11.0)
0.302
Events: 14; Global p-value (Log-Rank): 0.49321
AIC: 102.55; Concordance Index: 0.61
0.5
1
2
5
10

Supplementary Figure 1. A. Overall survival within NHL patients according to the EBV status (negative: red line; positive: green line) with adjustments made for patient age. **B.** Table representing a Cox regression model using tumor localization, age stratified into an old (>= 58 years old) and a young group (< 58 years old), EBV status, immune status stratified into immunocompetent (IC) group and immune-deficient group (ID: HIV + PTLD).



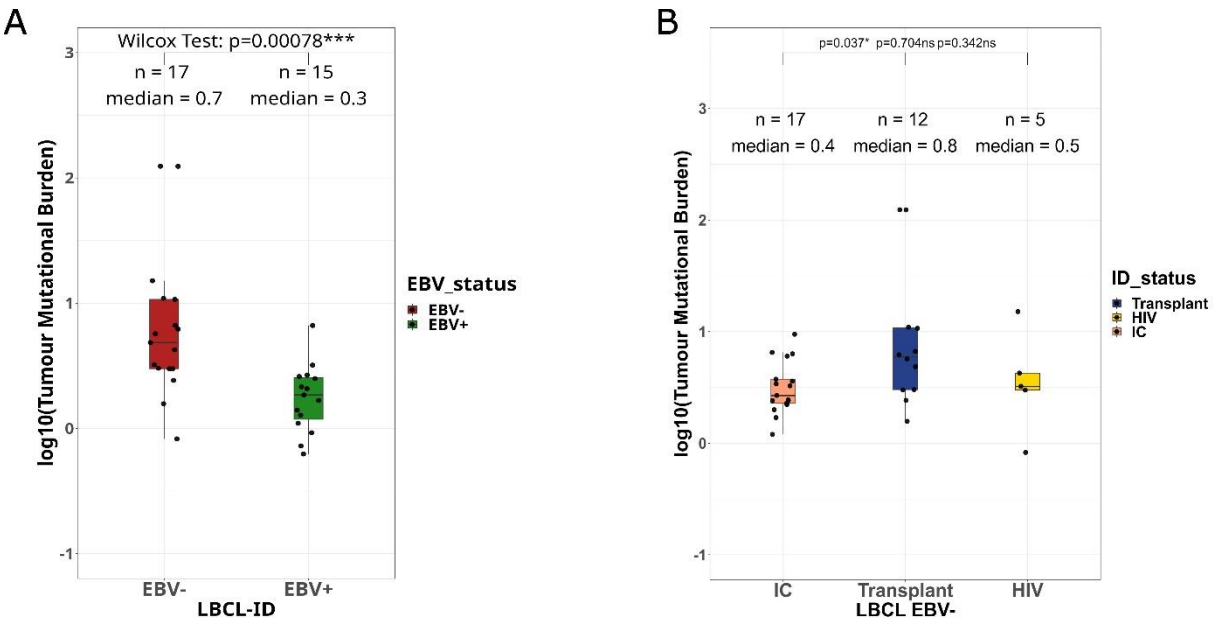
Supplementary Figure 2. A. Coverage of exome sequencing for tumors and their normal counterparts. Boxes are divided by median values. Length of boxes corresponds to interquartile

range and whiskers correspond to 1.5 interquartile ranges. **B**. RNA seq library size. **C.** Distribution of Log2 Count Per Million (CPM) across tumors.

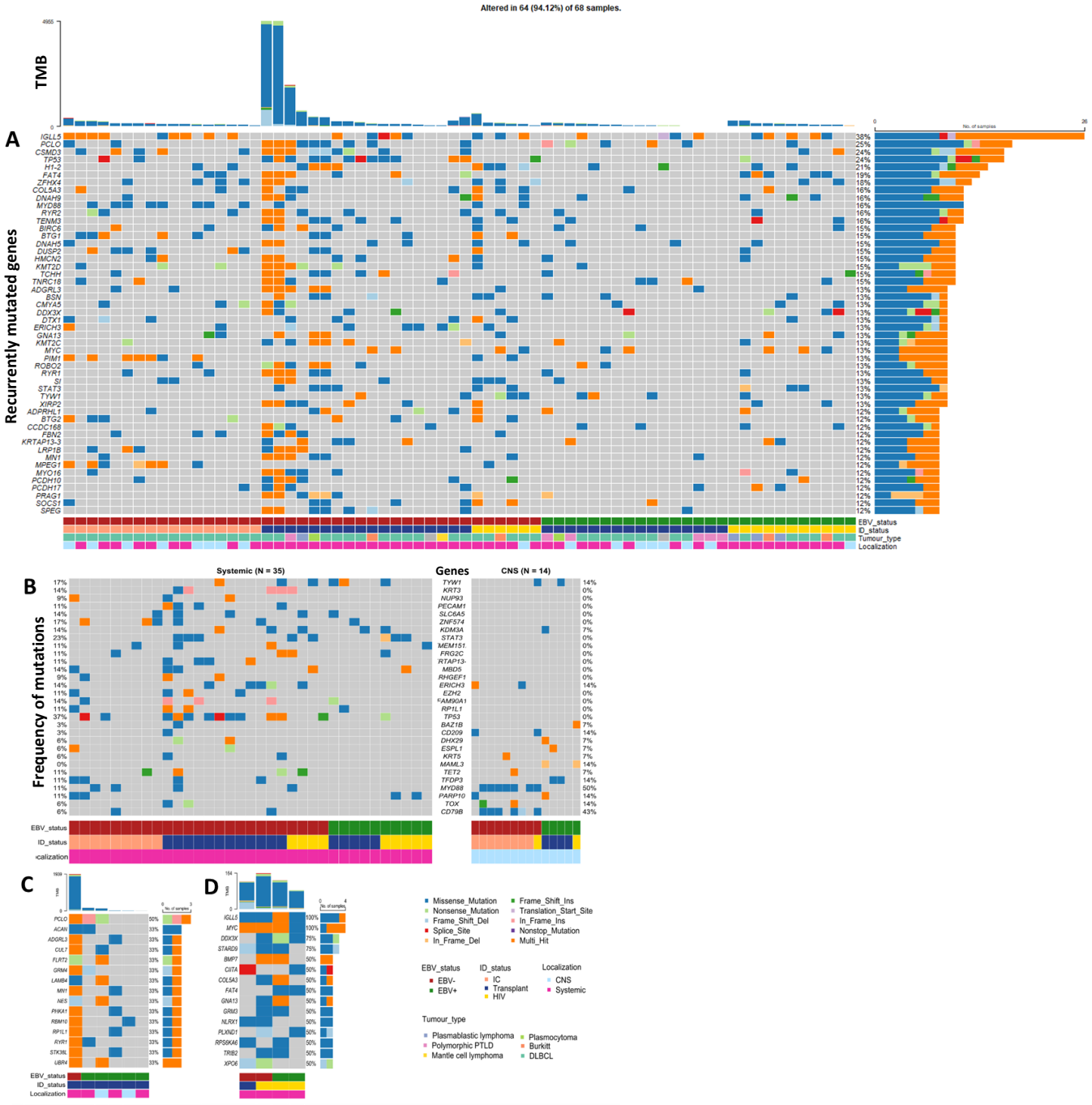


Supplementary Figure 3. **Tumor mutational burden**. The prevalence of somatic mutations in our cohort of NHL patients is illustrated among the ones of human cancer types described previously by [19]. The x axis shows the different cancers ordered based on their median numbers of somatic mutations. Our cohort is named "NHL" and appears close to the "DLBCL" one. The TMB is defined as the number of mutations per megabase. LAML: Acute Myeloid Leukemia, PCPG: Pheochromocytoma and Paraganglioma, THCA: Thyroid carcinoma, UVM: Uveal Melanoma, TGCT: Testicular Germ Cell Tumors, THYM: Thymoma, KICH: Kidney Chromophobe, ACC: Adrenocortical carcinoma, LGG: Brain Lower Grade Glioma, MESO: Mesothelioma, PRAD: Prostate adenocarcinoma, PAAD: Pancreatic adenocarcinoma, BRCA: Breast invasive carcinoma, SARC: Sarcoma, CHOL: Cholangiocarcinoma, UCS: Uterine Carcinosarcoma, GBM: Glioblastoma multiforme, KIRC: Kidney renal clear cell carcinoma, KIRP: Kidney renal papillary cell carcinoma, OV: Ovarian serous cystadenocarcinoma, UCEC: Uterine Corpus Endometrial Carcinoma, LIHC: Liver hepatocellular carcinoma, CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma, READ: Rectum adenocarcinoma, ESCA: Esophageal carcinoma, HNSC: Head and Neck squamous cell

carcinoma, DLBC: Lymphoid Neoplasm Diffuse Large B-cell Lymphoma, STAD: Stomach adenocarcinoma, COAD: Colon adenocarcinoma, NHL: Non-Hodgkin Lymphoma (IDEATION Cohort), BLCA: Bladder Urothelial Carcinoma, LUAD: Lung adenocarcinoma, LUSC: Lung squamous cell carcinoma, SKCM: Skin Cutaneous Melanoma.



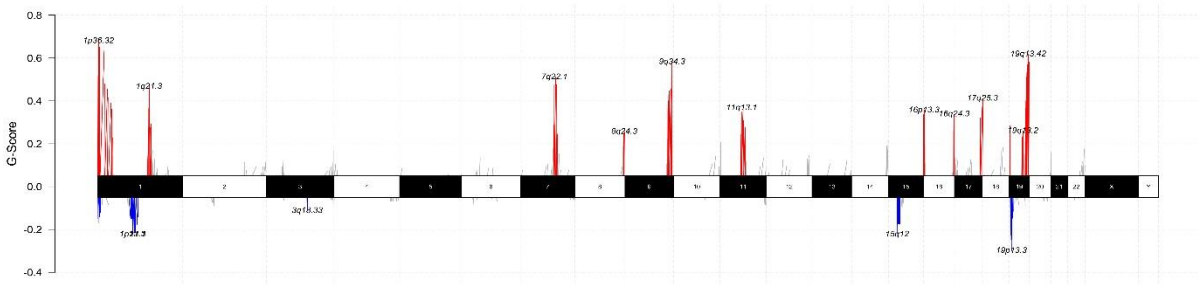
Supplementary Figure 4. Tumor mutational burden (TMB) according to **A**. the EBV status among the immunodeficient (ID) patients with large B cell lymphoma (LBCL) (n=32). **B**. and to the immune status among the EBV-negative LBCL patients (n=34): IC, transplant and HIV patients are illustrated with salmon, blue and HIV bars respectively. The TMB is defined as the number of mutations per megabase. IC: immunocompetent. *Wilcoxon Test*.



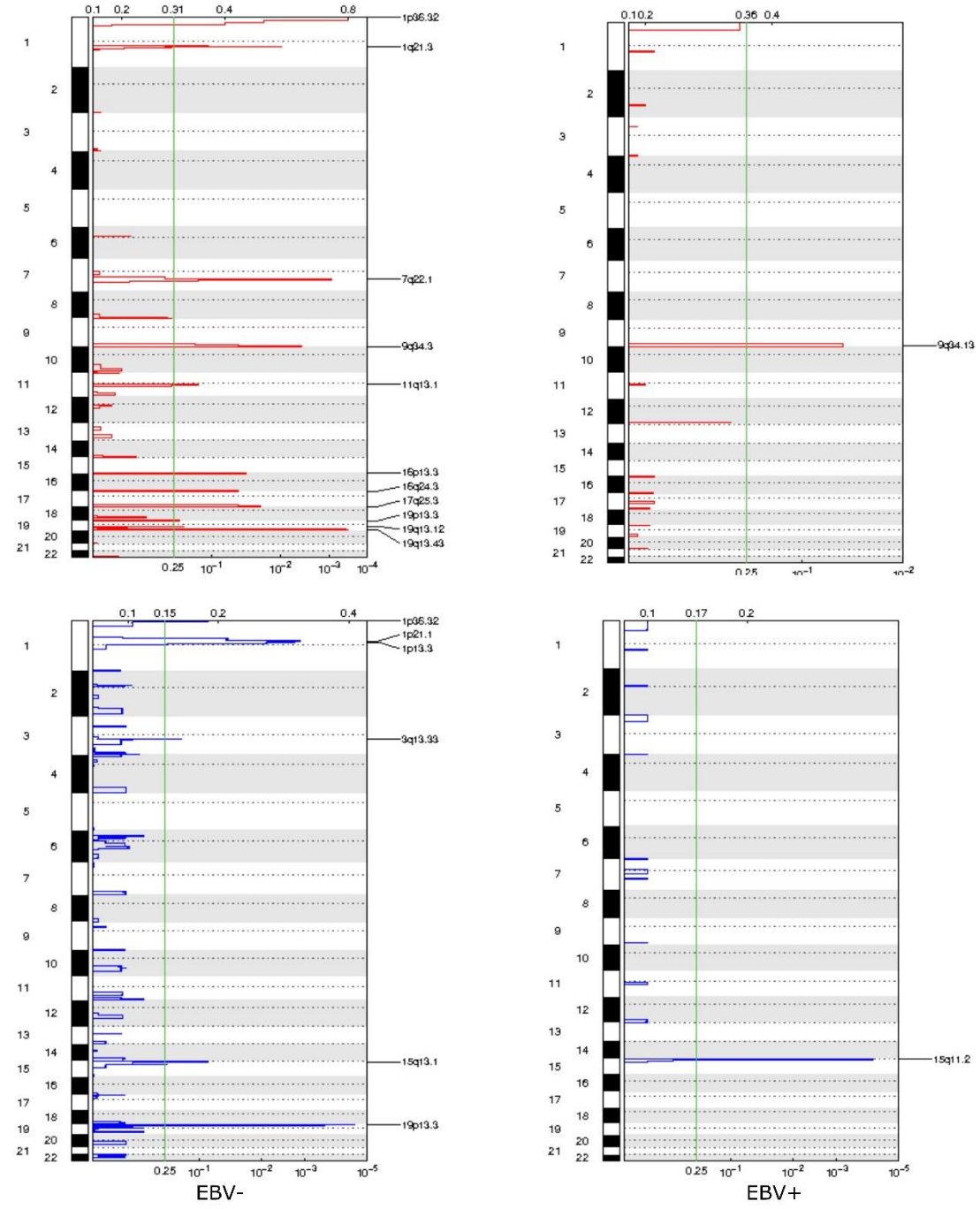
Supplementary Figure 5. **A**. Oncoplot of the top 50 most recurrently mutated genes within the overall population of 68 NHL patients. Tumor Mutational Burden (TMB) is represented as the total number of coding mutations per megabase per tumor. **B**. Co-oncoplot of the most recurrently mutated genes within the 49 LBCL samples with clinical annotations and according to the disease localization. **C**. Oncoplot of the most recurrently mutated genes within the 6 polymorphic PTLD with clinical annotations. **D.** Oncoplot of the most recurrently mutated

genes within the 4 Burkitt Lymphoma (BL) with clinical annotations. CNS: central nervous system. *Fisher exact test.*

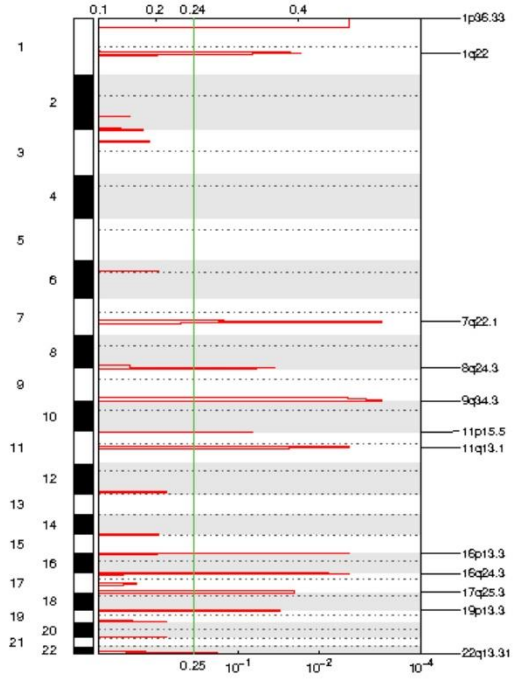
A

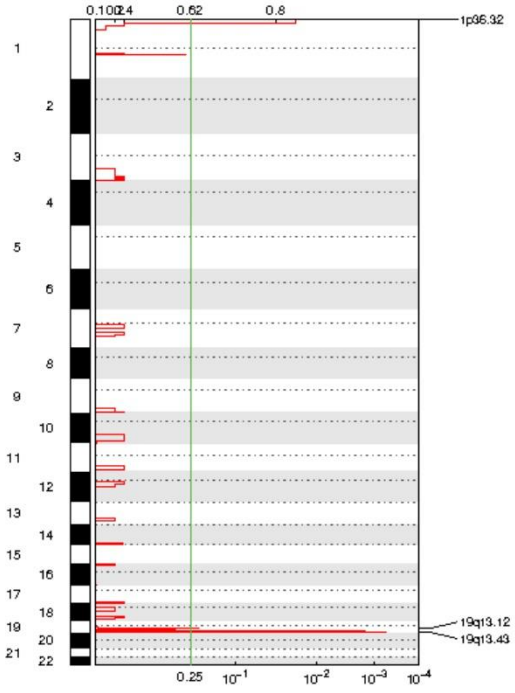

G-Score
1p36.32
1q21.3
7q22.1
9q34.3
9q34.3
11q13.1
16p13.3
16q24.3
17q25.3
19q13.42
19q13.2
3q13.33
15q11.2
19p13.3

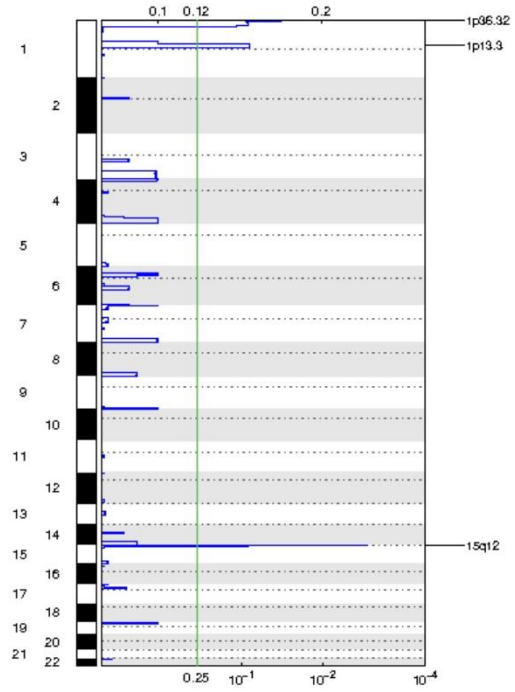
B


1p36.32
1q21.3
7q22.1
9q34.3
11q13.1
16p13.3
16q24.3
17q25.3
19p13.3
19q13.12
19q13.43
9q34.13
1p36.32
1p21.1
1p13.3
3q13.33
15q13.1
19p13.3
15q11.2
EBV-
EBV+

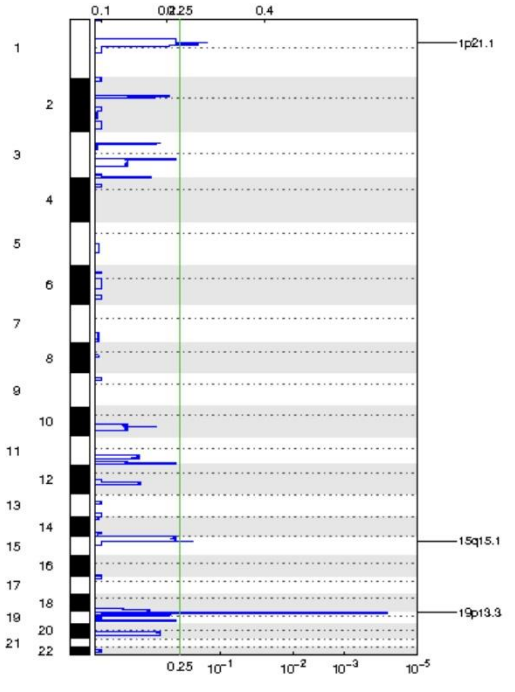
C


0.1
0.2
0.24
0.4
1p36.33
1q22
7q22.1
8q24.3
9q34.3
11p15.5
11q13.1
16p13.3
16q24.3
17q25.3
19p13.3
22q13.31
0.25
10-1
10-2
10-4


0.1002.4
0.62
0.8
1p36.32
19q13.12
19q13.43
0.25
10-1
10-2
10-3
10-4

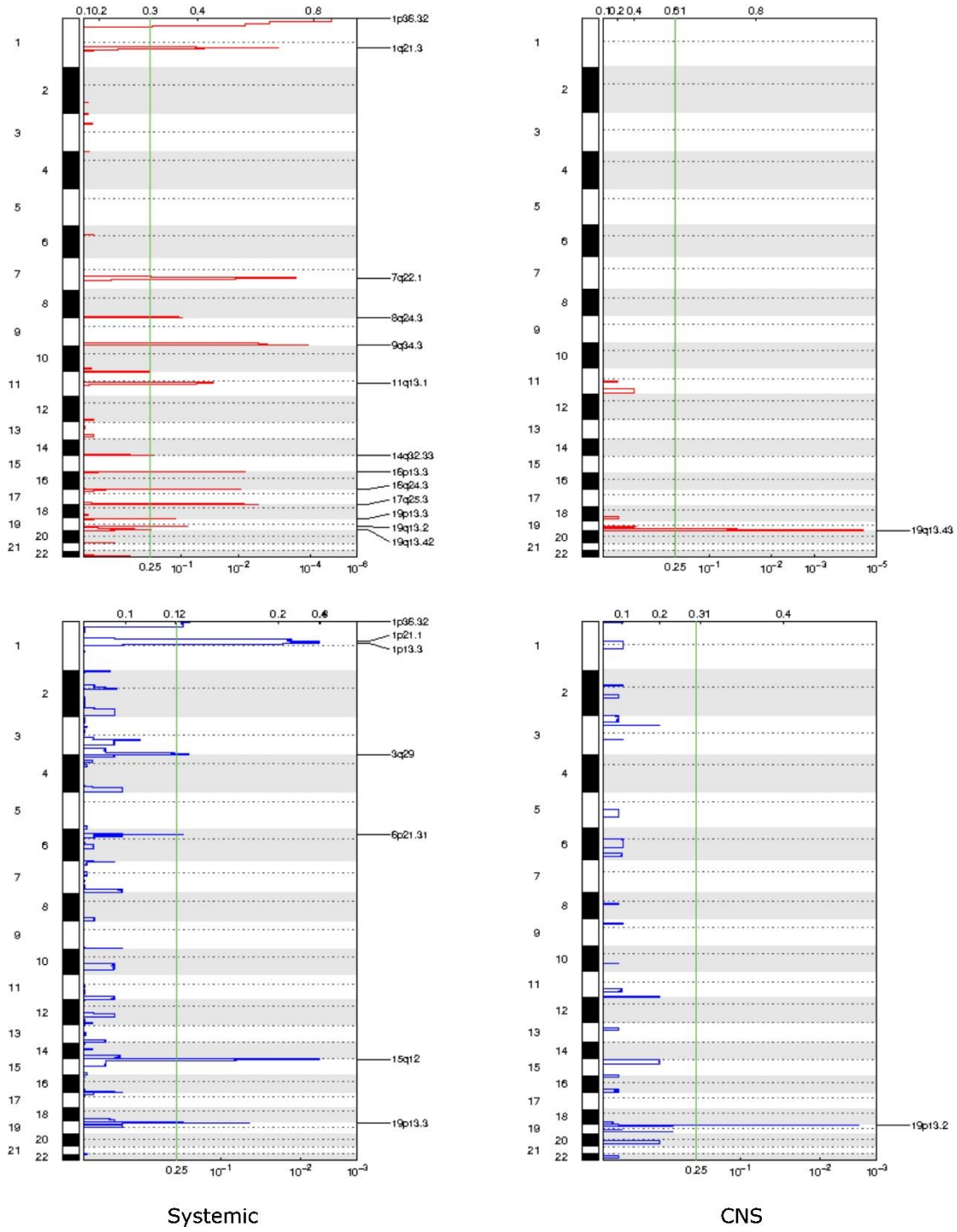

0.1
0.12
0.2
1p36.32
1p13.3
15q12
0.25
10-1
10-2
10-4

ID


0.1
0.25
0.4
1p21.1
15q15.1
19p13.3
0.25
10-1
10-2
10-3
10-5

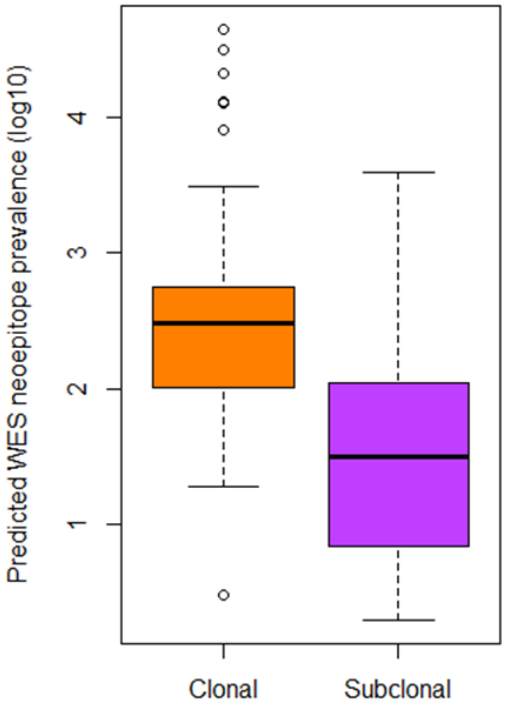
IC

D

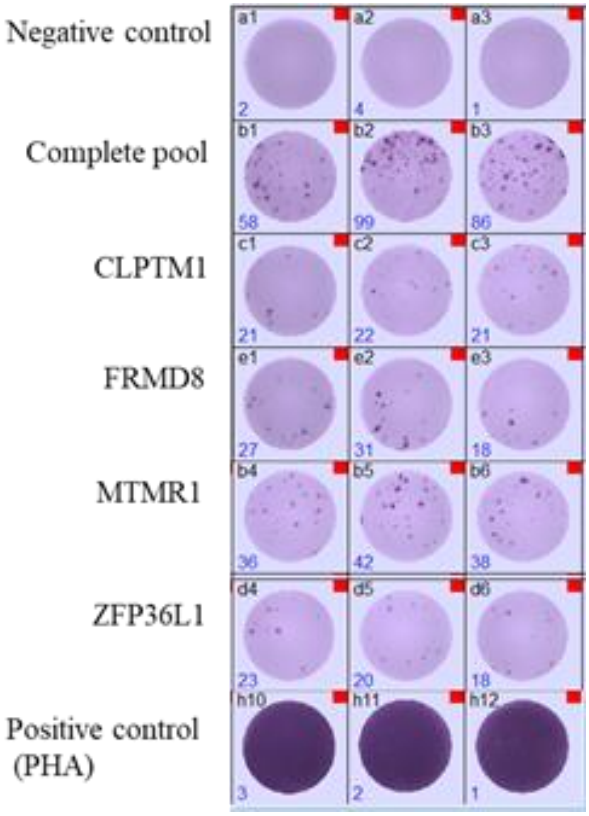


Supplementary Figure 6. **A**. **Frequency of arm-level somatic copy number alterations (SCNAs) among LBCL tumors.** dashed line indicating the threshold of arm-level SCNA frequency at 20% or higher. **B, C and D**. **Comparisons of GISTIC2.0-defined recurrent focal copy number** gains (red) and losses (blue). Chromosomes are shown on the

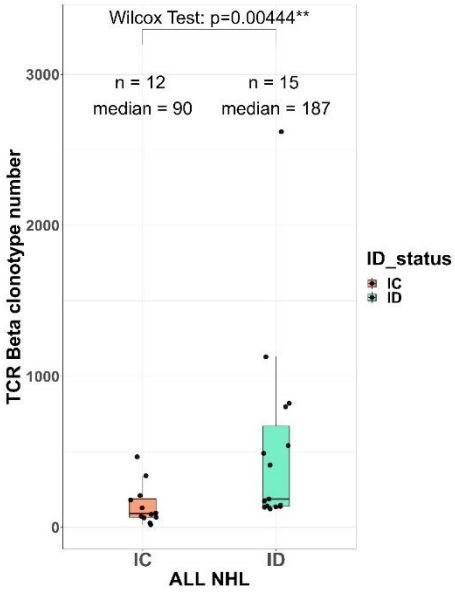
vertical axis. Green line denotes q-value of 0.25. Somatic copy number alterations (SCNAs) are labeled with their associated cytoband. **A.** Based on EBV status **B.** based on immune status. **C.** Based on disease localization. CNS: central nervous system; ID: immunocompetent; ID: immunodeficient.



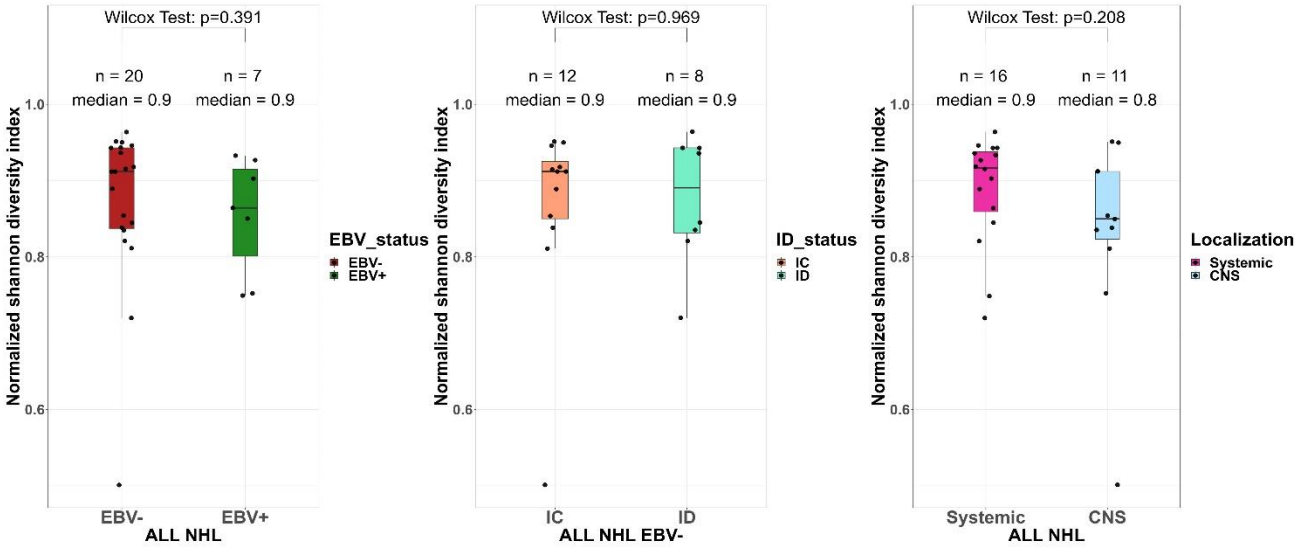
Supplementary Figure 7. Number of Clonal and Subclonal predicted neoepitopes (log_{10}) per patient based on WES sequencing.



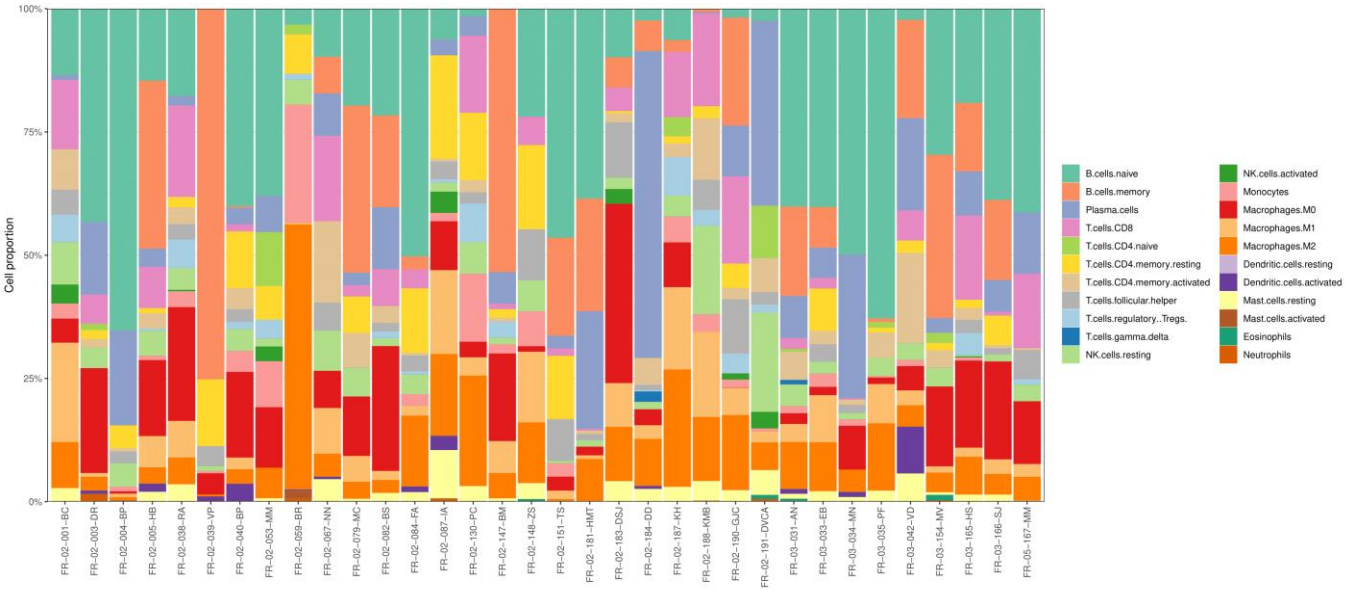
Supplementary Figure 8. Example ELISPOT results among one HIV patient with EBV-negative NHL: the patient was tested for 41 different peptides (only positive results against CLPTM1, FRMD8, MTMR1 and ZFP36L1 peptides are shown), negative control (at the top) and positive control (at the bottom). Thawed PBMC were co-cultured with personalized pooled peptides during 10 days and next tested for reactivity using IFN-γ enzyme linked immunoSpot (ELISPOT) assays. Patients were all tested for their personalized pooled peptides (named "complete pool") and eventually for each individual peptide if the number of cells were adequate (named as the mutated gene). The mean numbers of spot forming cells (SFC) from triplicates were normalized to number of IFN-γ spots detected per $1x10^6$ PBMC after background subtraction, and ELISPOT-IFN-γ positivity threshold was 50 SFC per million cells.



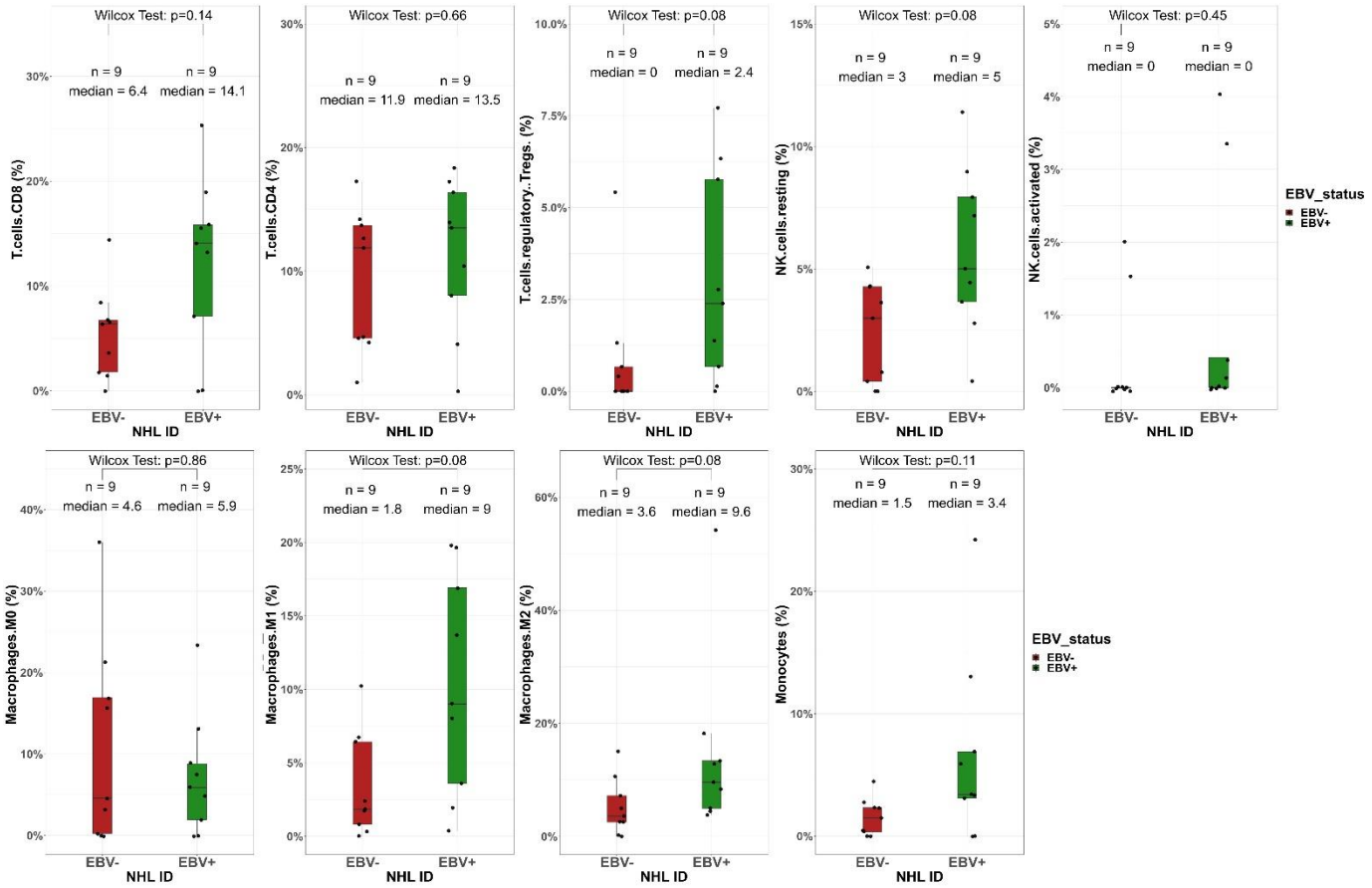
Supplementary Figure 9. Intra-tumoral TCR repertoire abundance according to the immune status within 27 NHL samples. *Wilcoxon Test*.



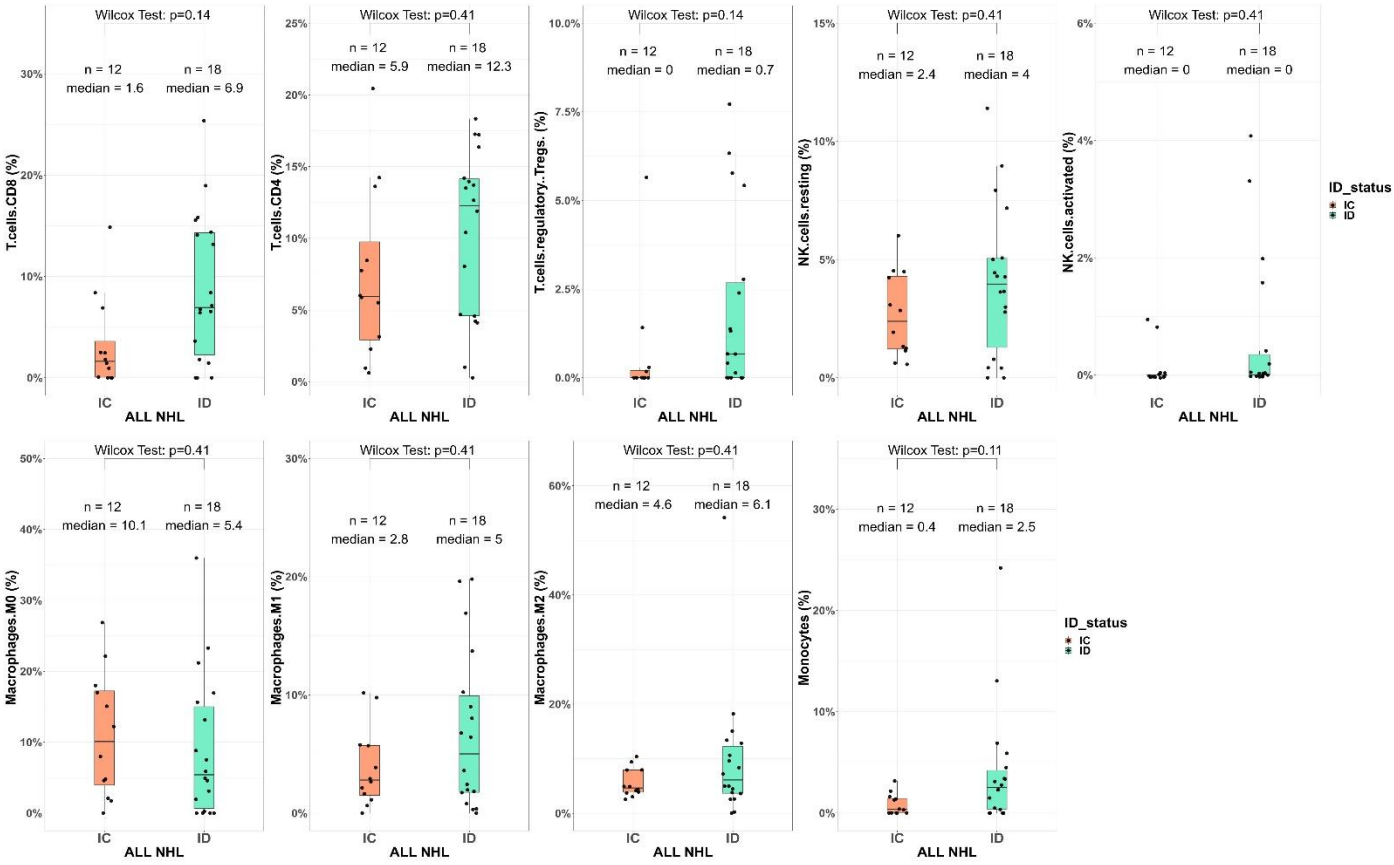
Supplementary Figure 10. Normalized Shannon index diversity index according to the EBV status (left), the immune status (middle) and the disease localization (right). Shannons' entropy was calculated using vdjtools V1.2.1 software[17] and was divided by (log (number of clonotypes)) to be independent of the abundance: lower values indicate more diversity. IC: immunocompetent; ID: immunodeficient; CNS: central nervous system. *Wilcoxon Test*



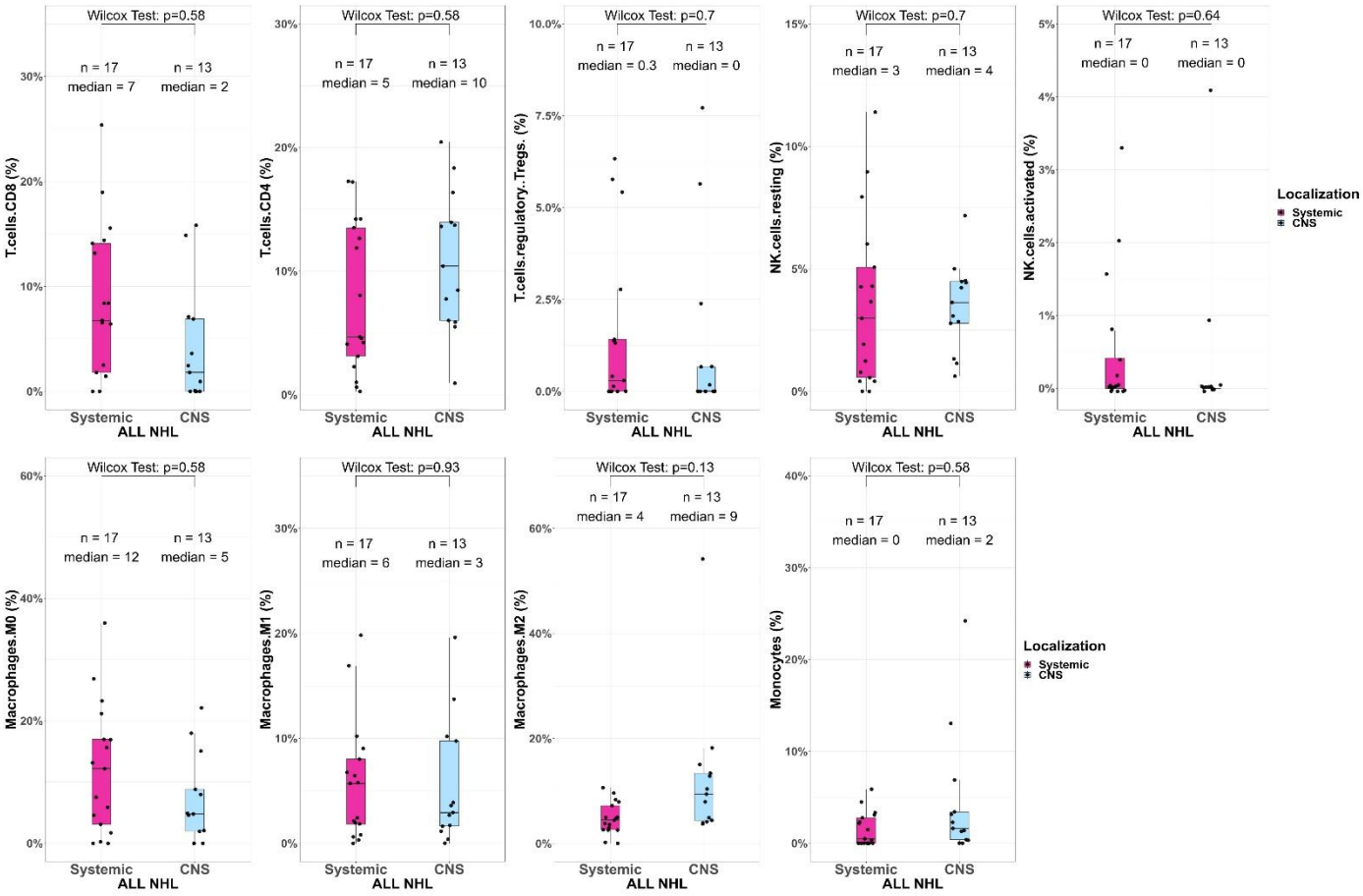
Supplementary Figure 11. Cell type abundance distribution for each patient using CIBERSORTx.



Supplementary Figure 12. Cell type abundance according to the EBV status among the ID patients only. Cell type abundance was assessed using CIBERSORTx.system. *Wilcoxon Test*.



Supplementary Figure 13. Cell type abundance according to the immune status. Cell type abundance was assessed using CIBERSORTx. IC: immunocompetent; ID: immunodeficient. *Wilcoxon Test*.



Supplementary Figure 14. Cell type abundance according to the disease localization. Cell type abundance was assessed using CIBERSORTx. CNS: central nervous system. *Wilcoxon Test*.

Tables

Supplementary Table 1. Comparison of the mutational frequencies stratified by EBV status. Mutational frequencies in the 34 EBV-negative lymphomas were compared to the 15 EBV-positive ones. Genes are ranked by uncorrected p values (with corrected p values shown in the adjacent column).

Hugo_Symbol	EBV-	EBV+	pval	OR	CI. low	CI. up	Adj.Pval
MYD88	11	0	0.01124097	Inf	Inf	1.352539871	0.3484702
HNRNPF	0	3	0.02469605	0.0000000	0.9941339	0.000000000	0.3827888
TP53	12	1	0.04263089	7.3914358	348.7086709	0.900912662	0.4405192
CD79B	8	0	0.08683169	Inf	Inf	0.831634188	0.4613095
AP3B2	0	2	0.08928571	0.0000000	2.2852181	0.000000000	0.4613095
MAML3	0	2	0.08928571	0.0000000	2.2852181	0.000000000	0.4613095
ESCO2	1	2	0.21846505	0.2046464	4.2426528	0.003250142	0.6768950
MYO15A	1	2	0.21846505	0.2046464	4.2426528	0.003250142	0.6768950
RTN4	1	2	0.21846505	0.2046464	4.2426528	0.003250142	0.6768950
TYW1	4	4	0.22720046	0.3752253	2.3835301	0.058448793	0.6768950
EZH2	4	0	0.29833959	Inf	Inf	0.291905276	0.6768950
KRTAP13-3	4	0	0.29833959	Inf	Inf	0.291905276	0.6768950
PECAM1	4	0	0.29833959	Inf	Inf	0.291905276	0.6768950
KRT3	5	0	0.30569453	Inf	Inf	0.412788484	0.6768950
KDM3A	3	3	0.35330099	0.3956178	3.3716730	0.046175510	0.6845207
PARP10	3	3	0.35330099	0.3956178	3.3716730	0.046175510	0.6845207
ARHGEF15	3	0	0.54325879	Inf	Inf	0.180556834	0.8872066
NUP93	3	0	0.54325879	Inf	Inf	0.180556834	0.8872066
TOX3	2	2	0.57635598	0.4147095	6.2831643	0.027324750	0.8872066
CEP250	3	2	0.63519543	0.6354177	8.4440603	0.064516010	0.8872066
PLXNB3	3	2	0.63519543	0.6354177	8.4440603	0.064516010	0.8872066
ZNF574	5	1	0.65177745	2.3768993	122.1232631	0.232289392	0.8872066
TNRC18	4	3	0.65987565	0.5407780	4.2502025	0.077917168	0.8872066
STAT3	5	3	0.68686965	0.6951469	5.1882401	0.113021613	0.8872066
DHX29	2	1	1.00000000	0.8774500	55.2479011	0.042342197	1.0000000
ESPL1	2	1	1.00000000	0.8774500	55.2479011	0.042342197	1.0000000
ID2	2	1	1.00000000	0.8774500	55.2479011	0.042342197	1.0000000
FOXQ1	5	2	1.00000000	1.1181454	13.2050937	0.156202537	1.0000000
GNAI2	4	1	1.00000000	1.8455237	98.3507357	0.161980906	1.0000000
RP1L1	3	1	1.00000000	1.3469268	76.1064554	0.097813307	1.0000000
TFDP3	4	2	1.00000000	0.8692637	10.7397691	0.108142054	1.0000000

Supplementary Table 2. Comparison of the mutational frequencies stratified by immune status. Mutational frequencies in the 32 ID LBCL were compared to the 15 IC DLBCL. Genes are ranked by uncorrected p values (with corrected p values shown in the adjacent column).

Hugo_Symbol	ID	IC	pval	OR	CI. up	CI. low	Adj. Pval
MYD88	3	8	0.004574082	0.1228121	0.644494	0.01724404	0.1326484
STAT3	8	0	0.038368700	Inf	Inf	1.02551071	0.3708974
TYW1	8	0	0.038368700	Inf	Inf	1.02551071	0.3708974
KDM3A	6	0	0.079848305	Inf	Inf	0.66525774	0.4791059
TP53	11	2	0.105125508	3.8314402	40.600568	0.68100497	0.4791059
ANO10	0	2	0.115646259	0.0000000	2.767552	0.00000000	0.4791059
PKDCC	0	2	0.115646259	0.0000000	2.767552	0.00000000	0.4791059
KRT3	5	0	0.148789334	Inf	Inf	0.50449696	0.5393613
E2F7	1	2	0.273122015	0.2496756	5.148406	0.00398254	0.6855032
RAPGEF6	1	2	0.273122015	0.2496756	5.148406	0.00398254	0.6855032
KRTAP13-3	4	0	0.283656478	Inf	Inf	0.35569410	0.6855032
PECAM1	4	0	0.283656478	Inf	Inf	0.35569410	0.6855032
PARP10	3	3	0.405458996	0.4905080	4.138715	0.05794602	0.8398793
ZNF574	3	3	0.405458996	0.4905080	4.138715	0.05794602	0.8398793
FOXD4L4	2	0	0.537414966	Inf	Inf	0.09910238	0.8924406
ARVCF	2	2	0.602031377	0.5076208	7.640914	0.03367628	0.8924406
SLAMF7	2	2	0.602031377	0.5076208	7.640914	0.03367628	0.8924406
SPTAN1	2	2	0.602031377	0.5076208	7.640914	0.03367628	0.8924406
FAM90A1	4	1	0.646250113	2.2520240	119.418697	0.19933525	0.8924406
MBD5	4	1	0.646250113	2.2520240	119.418697	0.19933525	0.8924406
SLC6A5	4	1	0.646250113	2.2520240	119.418697	0.19933525	0.8924406
ADCY8	4	2	1.000000000	1.0699424	13.119541	0.13457850	1.0000000
ERICH3	5	2	1.000000000	1.3799652	16.160751	0.19512568	1.0000000
ESPL1	2	1	1.000000000	1.0652809	66.787013	0.05166123	1.0000000
FRG2C	3	1	1.000000000	1.6394834	92.199004	0.11990355	1.0000000
MCAT	3	2	1.000000000	0.7800302	10.287934	0.07993619	1.0000000
NUP93	2	1	1.000000000	1.0652809	66.787013	0.05166123	1.0000000
RP1L1	3	1	1.000000000	1.6394834	92.199004	0.11990355	1.0000000
TFDP3	4	2	1.000000000	1.0699424	13.119541	0.13457850	1.0000000

IC: immunocompetent; ID: immunodeficient; Adj. P value: adjusted P value after FDR correction.

Supplementary Table 3 is available in a separate excel spreadsheet. All GISTIC2 thresholded SCNA by gene.

References

1. Swerdlow SH, Campo E, Pileri SA, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*. 2016;127(20):2375–2390.
2. Chen X, Chang C-W, Spoerke JM, et al. Low-pass Whole-genome Sequencing of Circulating Cell-free DNA Demonstrates Dynamic Changes in Genomic Copy Number in a Squamous Lung Cancer Clinical Cohort. *Clin Cancer Res*. 2019;25(7):2254–2263.
3. Adalsteinsson VA, Ha G, Freeman SS, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun*. 2017;8(1):1324.
4. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213–219.
5. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*. 2018;15(8):591–594.
6. Costello M, Pugh TJ, Fennell TJ, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res*. 2013;41(6):e67.
7. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
8. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*. 2013;29(18):2238–2244.
9. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res*. 2016;44(16):e131.
10. Mermel CH, Schumacher SE, Hill B, 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):R41.
11. Gillis S, Roth A. PyClone-VI: scalable inference of clonal population structures using whole genome data. *BMC Bioinformatics*. 2020;21(1):571.
12. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
13. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.
14. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25.
15. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
16. Bolotin DA, Poslavsky S, Mitrophanov I, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods*. 2015;12(5):380–381.
17. Shugay M, Bagaev DV, Turchaninova MA, et al. VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLoS Comput Biol*. 2015;11(11):e1004503.
18. Goncharov M, Bagaev D, Shcherbinin D, et al. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nat Methods*. 2022;19(9):1017–1019.
19. Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415–421.