Multi-Omics Analysis of Primary Central Nervous System Lymphoma

Isaias Hernández-Verdin

Thèse de doctorat

NNT : 2022UPASL028

université PARIS-SACLAY

Multi-omics analysis of primary central nervous system lymphoma

Analyse multi-omique des lymphomes primitifs du système nerveux central

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°582 Cancérologie, biologie, médecine, santé (CBMS)
Spécialité de doctorat : Sciences de la vie et de la santé
Graduate School : Sciences de la vie et santé
Référent : Faculté de médecine

Thèse préparée dans l'unité de recherche **Institut du cerveau** (AP-HP, CNRS, Inserm, Sorbonne Université) sous la direction du Pr **Khê HOANG-XUAN**, PU-PH, HDR, Sorbonne Université, et le co-encadrement du Dr **Agusti ALENTORN**, MD, PhD, Sorbonne Université

Thèse présentée et soutenue à Paris, le 06 mai 2022, par

Isaias HERNANDEZ-VERDIN

Composition du jury :

Pr THIEBLEMONT Catherine PU-PH, Université Paris Cité	Présidente
Pr JARDIN Fabrice PU-PH, Centre Henri Becquerel, Université de Rouen	Rapporteur et Examinateur
Pr LECHAPT-ZALCMAN Emmanuèle PU-PH, Université Paris-Est Créteil	Rapporteuse et Examinatrice
Dr CAVALLI Florence CRCN, Institut Curie	Examinatrice
Pr DESESTRET Virginie PU-PH, Université Claude Bernard Lyon 1, Université de Lyon	Examinatrice
Pr HOANG-XUAN Khê PU-PH, Sorbonne Université	Directeur de thèse

Acknowledgements

The results imprinted in this thesis were not only the product of my dedication and work but also of the support, advices, patience, and knowledge of many others. Firstly, I express my deep gratitude to Monica Sierra del Rio, a friend and collaborator, who contacted me with my current thesis supervisors. My profound appreciation goes to Pr. Khê HOANG and Dr. Agusti ALENTORN for having faith in me along with patience, support, encouragement, knowledge, and guidance throughout this process. Thanks to Pr. Marc SANSON and Dr. Emmanuelle HUILLARD for welcoming me to the "Genetics and Development of Nervous System Tumors" at Paris Brain Institute.

Thanks to all my colleagues at Paris Brain Institute, specially in the Sanson-Huillard team, for all the conversations, work feedbacks, lunchs, bars, and moments together. I am grateful to the "Association pour la Recherche sur les Tumeurs Cérébrales" (A.R.T.C) and to "Assistance Publique - Hôpitaux de Paris" (AP-HP) for funding me during all my thesis.

I would also like to extend my appreciation to "La Ligue contre le cancer" through the program "Cartes d'Identité des Tumeurs" along with all the involved French hospitals, clinicians, technicians, etc, for making the project possible. Thanks to Yannick Marie and the sequencing platform at the Paris Brain Institute and all people at Mazarin for helping with sequencing or clinical data.

To my friends and colleagues in the lab, Irma, Yanis, Quentin, Mohamed, Alexa, Julie, Diego, Alberto, Fernando for your friendship, support, and all the incredible moments we shared throughout this journey.

Thanks to all my family, especially to my parents for all the support and love despite the long distance. My deepest gratitude to my beloved Jimena, for all her support at ups and downs and love. It is practically impossible to summarize all I learned, both academically and personally, in the past three years; however, I am grateful for all the single things I learned from each one of you.

Thanks

Table of Contents

List of Abbreviations

5mc cytosine to 5-methylcytosine. 45

ABC activated B-cell–like. 11

ACC adenoid cystic carcinoma. 96

AdenoCa adenocarcinoma. 98

AID activation-induced cytidine deaminase. 9

APC antigen presenting cell. 21

APM antigen processing and presentation machinery. 22

ARTC Association pour la Recherche sur les Tumeurs Cérébrales. 84

ASCT autologous hematopoietic stem cell transplantation. 29

ASHM aberrant somatic hypermutation. 13

AUROC area under the ROC curve. 61

B2M beta-2-microglobulin. 22

BAFF-R B-cell activating factor receptor. 7

BAFs B-allele fractions. 41

BBB blood-brain barrier. 18

BCC basal cell carcinoma. 95

BCR B-cell receptor. 7

BER base excision repair. 9

BL Burkitt lymphoma. 11

BLCA bladder cancer. 97

BLIMP1 B-lymphocyte-induced maturation protein 1. 9

BM bone marrow. 6

CESC cervical squamous cell carcinoma. 97

C-index concordance index. 60

CAFs Cancer-associated fibroblasts. 22

c-AID canonical-AID. 9

CART-cell chimeric antigen receptor T-cell. 35

CBM CARD11–BCL10–MALT1. 14

CDK cyclin-dependent kinase. 35

CI confidence interval. 60

CLL chronic lymphoid leukaemia. 43

CNAs copy-number alterations. 15

CNS central nervous system. ii, 1

COO cell of origin. 11

COSMIC Catalogue Of Somatic Mutations In Cancer. 9

CoxPH Cox proportional hazards. 57

CPI Cluster Prediction Index. 62

CR Complete remission. 33

CS significant clusters. 126

CSF cerebrospinal fluid. 4

CSR class-switch recombination. 9

CTLs cytotoxic T-lymphocytes. 18

CYT cytolytic index. 54

DCs dendritic cells. 18

DEGs differentially expressed genes. 49

dES differential enrichment scores. 52

DLBCL diffuse large B-cell lymphoma. 1

DM Differential methylation. 47

DNA deoxyribonucleic acid. 9

DNMTs DNA methyltransferase enzymes. 45

DPI Data Processing Inequality. 51

DZ dark zone. 9

EBV Epstein-Barr virus. 1

ECM extracellular matrix. 18

EMZL extranodal MZL of mucosa-associated lymphoid tissue. 82

epiCMIT epigenetically-determined Cumulative MIToses. 13

FDA Food and Drug Administration. 23

FDC follicular dendritic cells. 9

FF fresh-frozen. 4

FFPE formalin-fixed, paraffin-embedded. 4

FL follicular lymphoma. 11

FRCs fibroblastic reticullar cells. 22

GATK Genome Analysis Toolkit. 40

GC germinal-center. 4

GCB germinal center B-cell–like. 11

GEP gene expression profiling. 11

GO Gene Ontology. 50

GSA gene set analysis. 50

GSEA Gene Set Enrichment Analysis. 50

GSVA Gene Set Variation Analysis. 50

GWAS genome-wide association studies. 68, 73

HAART Highly Active Antiretroviral Therapy. 2

HCL human cell landscape. 94

HD high-dose. 29

HIV human immunodeficiency virus. 2

HL Hodgkin's lymphoma. 75

HNSC non-small cell lung cancer. 95

HPV human papillomavirus. 97

HR hazard ratio. 60

HSCs hematopoietic stem cells. 6

IC immune checkpoint. 23

ICIs immune checkpoint inhibitors. 23

IELSG International Extranodal Lymphoma Study Group. 32

IFN-γ interferon-γ. 19

IG immunoglobulin. 6

IHC Immunohistochemistry. 4

indels insertions/deletions. 39

intNMF integrative non-negative matrix factorization. 64

IPI international prognostic index. 25

KEGG Kyoto Encyclopedia of genes and genomes. 50

Kidney-RCC kidney renal cell carcinoma. 98

KIRC kidney renal clear cell carcinoma. 96

KIRP kidney renal papillary cell carcinoma. 96

KM Kaplan-Meier. 25, 57

KPS Karnofsky performance score. 32

LASSO least absolute shrinkage and selection operator. 64

LECs lymphatic endothelial cells. 25, 27

LIHC liver hepatocellular carcinoma. 96

LOLA Locus Overlap enrichment Analysis. 47

LUAD lung adenocarcinoma. 96

LZ light zone. 9

mad median absolute deviation. 61

MAF mutation annotated format. 40

MB medulloblastoma. 95

MCA mouse cell atlas. 94

MCL mantle cell lymphoma. 11

MDSCs myeloid-derived suppressor cells. 19

MHC major histocompatibility complex. 4

MI mutual information. 51

MM multiple myeloma. 11

MMP9 matrix metalloproteinase-9. 22

MMR mismatch repair. 9

MPPs multipotent progenitors. 6

MRI magnetic resonance imaging. 3

MRP multidrug resistance protein. 79

MRs master regulators. 51

MS multiple sclerosis. 77

MSKCC Memorial Sloan-Kettering Cancer Center. 32

MTX methrotexate. 29

MZL marginal zone lymphoma. 11

NCCN National Comprehensive Cancer Network. 33

NESs normalized enrichment scores. 51

NF-$\kappa\beta$ Nuclear Factor-$\kappa\beta$. 14

NGS next-generation sequencing. 12

NHEJ nonhomologous end joining. 95

NHLs non-Hodgkin lymphomas. 1

NK natural killer. 6

NMF nonnegative matrix factorization. 43

NMZL nodal MZL. 82

NSCLC non-small cell lung cancer. 95

OS overall survival. 3

OV ovarian cancer. 96

Panc-Endocrine pancreatic neuroendocrine tumors. 98

PBMCs peripheral blood mononuclear cells. 94

PCA principal component analysis. 61

PCNSL Primary central nervous system lymphoma. ii, 1, 2, 4

PFS progression-free survival. 25

pHLA peptide-HLA-I. 21

PINSPlus Perturbation clustering for data INtegration and disease Subtyping. 64

PON panel of normals. 40

R/R refractory/relapsed. 32

RA rheumatoid arthritis. 77

RBraLymP RNA-based Brain Lymphoma Profiler. 133

RNA-seq RNA sequencing. 48

SAM Sequence alignment/map. 39

SBS single base substitution. 43

SCC squamous cell carcinoma. 98

scRNA-seq single-cell RNA-sequencing. 13

sd standard deviation. 61

SHM somatic hypermutation. 9

SKCM skin cutaneous melanoma. 95

SLE systemic lupus erythematosus. 77

SMZL splenic MZL. 82

SNVs single nucleotide variants. 39

SOM self-organizing map. 66

ssGSEA single sample Gene Set Enrichment Analysis. 50

STAD stomach adenocarcinoma. 97

SVM-RFE support vector machine-recursive feature elimination. 62

SVs structural variants. 15

TADs topologically associated domains. 98

TAMs tumor-associated macrophages. 19

TAP transporter associated with antigen processing. 22

TCC transitional cell carcinoma. 98

TCR T-cell receptor. 18

TFBS Transcription factor binding sites. 47

TFH follicular helper T-cells. 9

TFs transcription factors. 6

TH T-helper. 9

THCA thyroid cancer. 96

Thy-AdenoCA thyroid adenocarcinoma. 98

TILs tumor-infiltrating lymphocytes. 23

TLR toll-like receptor. 14

TMB tumor mutational burden. 19

TME tumor microenvironment. 13, 15

TMZ temozolomide. 35

Tregs T-regulatory cells. 19

TSGs tumor suppressor genes. 46

TSS transcription start sites. 99

USA United States of America. 2

VCF variant call format. 40

VECs vascular endothelial cells. 22

VEP Variant Effect Predictor. 40

VIPER virtual inference of protein activity by enriched regulon analysis. 51

WBRT whole-brain radiotherapy. 29

WES whole-exome sequencing. 39

WGBS whole-genome bisulfite sequencing. 10

WGS whole-genome sequencing. 39

WHO World Health Organization. 1

List of Figures

List of Tables

Introduction

Chapter 1

Overview of PCNSL

1.1 PCNSL in the context of large B-cell lymphomas and CNS tumors

The definition of primary CNS lymphoma has evolved along with the development of modern morphological, immunological, and molecular cytogenetic techniques. The most recent revised World Health Organization (WHO) classification of Tumours of Haematopoietic and Lymphoid Tissues (Steven H. Swerdlow et al., 2008; Steven H. Swerdlow et al., 2016) identifies PCNSL as a separate entity defined as diffuse large B-cell lymphoma (DLBCL) arising exclusively within the brain, spinal cord, leptomeninges, or eye. This excludes other lymphoma entities involving the CNS such as intravascular large B-cell lymphomas, low grade lymphomas, T cell CNS lymphomas, Burkitt's lymphoma, ALK lymphoma of the CNS, and CNS lymphomatoid granulomatosis. Moreover, immunodeficiency-associated PCNSL, normally associated to Epstein-Barr virus (EBV), has been recently proven to be a rare PCNSL subtype ($<$ 10% of cases) (Gandhi et al., 2021).

DLBCL is defined as a neoplasm of medium or large B lymphoid cells whose nuclei are the same size as, or larger than, those of normal macrophages, or more than twice the size of those of normal lymphocytes, with a diffuse growth pattern (Steven H. Swerdlow et al., 2008).

While DLBCL constitutes 25—35% of adult non-Hodgkin lymphomas (NHLs), PCNSL is estimated to account for up to 1-2% of NHLs, 4-6% of all extranodal lymphomas, and about 2-3% of all CNS tumors (Hoang-Xuan et al., 2015; Teras et al., 2016) (Figure 1.1). If immunodeficiency is a well known risk factor, PCNSL

in immunocompromised patients has become rare. Indeed, the incidence of AIDS related PCNSL has been dramatically reduced since the introduction of Highly Active Antiretroviral Therapy (HAART) in the late 90's. The prevalence of PCNSL is less than 10% in human immunodeficiency virus (HIV) patients (Gandhi et al., 2021). Today, patients receiving prolonged immunosuppressive agents represent the major at risk population (solid organ transplantation, autoimmune disorders) (Franca et al., 2020).

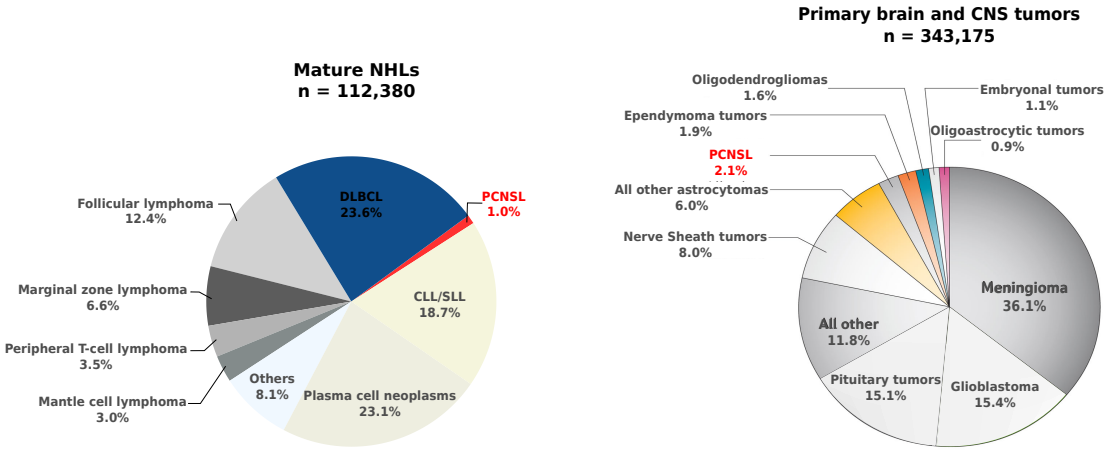


Figure 1.1: Relative frequency of lymphomas and of primary brain and central nervous system tumours. The left figure shows the NHLs neoplasms estimated cases from the North American Association of Central Cancer Registries (NAACCR, from 1995 to 2016); while the right figure shows the Central Brain Tumour Registry of the United States (CBTRUS) statistical report by histological groupings (n = 343,175). Adapted from Teras et al., 2016, Ostrom et al., 2014, and Weller, et al., 2015.

Despite being regarded for a long time as DLBCL in the CNS, mainly because it shows histology of DLBCL, PCNSL has been proved to be molecularly a different biological entity (Yoshida et al., 2016). As PCNSL in the immunocompetent population represent the vast majority of the patients, with a lymphomagenesis distinct from that related to immunodeficient PCNSL (involving EBV), my thesis is focused on PCNSL in immunocompetent patients.

1.2 Epidemiology of PCNSL

The incidence of PCNSL in France and in the United States of America (USA) is very similar, being 0.45 and 0.44 per 100,000 persons in the periods of 1990-2006 and 2009-2015, respectively. Nevertheless, the incidence rate in the United states (based on the SEER registry database) has grown continuously from 1975 to 2017 with a 5 fold increase, reaching an incidence of up to 4.32/100 000 in patients aged 70-79 years (reported by the Central Brain Tumor Registry) (Eloranta et al., 2018; Farrall & Smith, 2021; Haldorsen et al., 2007; Lv et al., 2022; Makino, Nakamura,

Kino, Takeshima, & Kuratsu, 2006; Mendez et al., 2018; Meulen, Dinmohamed, Visser, Doorduijn, & Bromberg, 2017; Ostrom et al., 2017; Shiels et al., 2016; Shin et al., 2015; Villano, Koshy, Shaikh, Dolecek, & McCarthy, 2011). This lymphoma can be presented at any age but has a peak incidence at 65 years with a male-to-female ratio of 1.5 in both France and the USA (Eloranta et al., 2018; Le Guyader-Peyrou et al., 2019; Ostrom et al., 2017; Steven H. Swerdlow et al., 2008). When removing the immunodeficient PCNSL patients, the overall rates of PCNSL in the immunocompetent population were stable during 1992–2011, on the contrary, rates among patients aged over 65 increased regardless of gender about 1.7% per year (Shiels et al., 2016). The median age at diagnosis is approximately 67 years in recent large cohorts while the prevalence of patients older than 60 years is between 60 and 70% (Eloranta et al., 2018; Farrall & Smith, 2021; Houillier et al., 2020; Meulen, Dinmohamed, Visser, Doorduijn, & Bromberg, 2017; Shin et al., 2015).

Prognosis, of PCNSL is significantly worse than that of DLBCL, with a median overall survival (OS) and 5 year survival of 26 months and 22% for PCNSL versus 124 months and 51% for systemic DLBCL (Horvat et al., 2018; Houillier et al., 2020; Yoshida et al., 2016). However there is a wide heterogeneity within PCNSL and the underlying physiopathological reasons of the clinical behavior of the tumors are not yet elucidated.

1.3 PCNSL clinical presentation and diagnosis

PCNSL more frequent symptoms include cognitive dysfunction, psychomotor slowing, focal neurological deficits, increased intracranial pressure, personality changes, and weakness, which might be explained by lymphoma cells infiltration into the white matter tracts of the corpus callosum and internal capsule. Seizures are less frequent (10%) compared to gliomas (Hoang-Xuan et al., 2015; Houillier et al., 2020; Shao et al., 2021; Steven H. Swerdlow et al., 2008). Additionally, when a PCNSL is found it is rarely the result of a systemic lymphoma's metastasis (only 5% of cases) (Houillier et al., 2020).

Diagnosis starts with magnetic resonance imaging (MRI) which typically shows hypointense on T1-weighted and isointense to hyperintense on T2-weighted images and enhancing single lesions (70%) or multiple lesions (30%) with modest surrounding edema, usually located in periventricular areas and/or deep gray matter (Figure 1.2). Meningeal involvement may present as foci of abnormal contrast enhancement. Furthermore, the use of corticosteroids is highly discouraged since it has been proved to make lesions vanish within hours, hence the term ghost tumors (Houillier et al., 2020). Regarding the localization, about 60% of PCNSLs are supratentorial lesions happening in the frontal lobe (15% of cases), temporal lobe (8%), parietal lobe (7%), occipital lobe (3%), basal ganglia, and periventricular brain parenchyma (10%), corpus callosum (5%), posterior fossa (13%), and spinal cord (in 1%) (Steven H. Swerdlow et al., 2008).

MRI findings, although suggestive, are not specific enough and need to be com-

plemented with pathological confirmation which in most cases relies on stereotactic needle biopsy in the absence of corticosteroids. The biopsy can only be avoided when lymphoma cells are discovered in the cerebrospinal fluid (CSF) (10–30%) or a vitreous-body biopsy (uveitis found by slit-lamp examination in 10–20% of cases).

Pathologically, PCNSL cells present frequent perivascular infiltration and consist, cytomorphologically speaking, of atypical cells with medium-sized to large round, oval, irregular, or pleomorphic nuclei and distinct nucleoli, corresponding to centroblasts or immunoblasts (Figure 1.2) (Ricard et al., 2012; Steven H. Swerdlow et al., 2008).

Immunohistochemistry (IHC) helps histology to yield the diagnosis of PCNSL where almost all PCNSL cells express pan-B-cell markers like CD20, CD19, CD22, CD79A, IgM, and IgD but not IgG. B-cell differentiation markers, BCL-6 (60–80%) for germinal-center (GC) B-cells, and IRF4/MUM1 (90%) for late GC B-cells and plasma cells, are also important for PCNSL diagnosis. PCNSL, hence, has been defined immunophenotypically as post-GC B-cells. Moreover, cells have a very high proliferative activity since the Ki-67 proliferation index is usually >70% and can even be >90%. Surprisingly, even though the loss of the major histocompatibility complex (MHC) happens in approximately 50% of PCNSL, HLA-A, HLA-B, HLA-C, and HLA-DR are variably expressed (Figure 1.2) (Baumgarten et al., 2018; Steven H. Swerdlow et al., 2008). Of note, these IHC analyses are performed on formalin-fixed, paraffin-embedded (FFPE) tissues but, for research purposes, collecting fresh-frozen (FF) samples and/or developing strategies to account for the genetic material degradation due to the fixation process, should be encouraged.

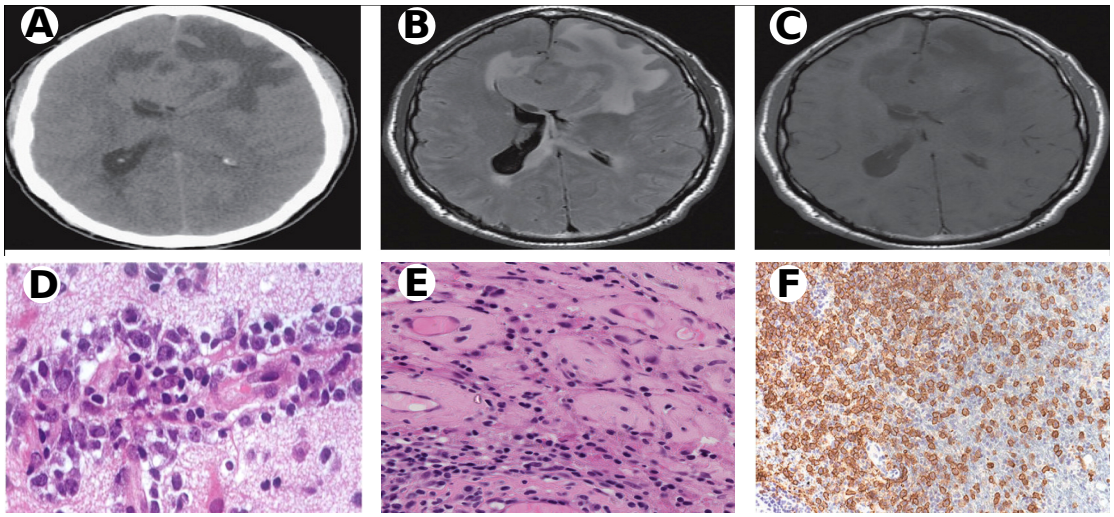


Figure 1.2: PCNSL diagnosis by MRI, histopathology and IHC. Computed tomography scan showing mild hyperdense lesion (Panel A). Peripheral oedema shown by a fluid attenuated inversion recovery (FLAIR)-weighted (Panel B) and by a T1-weighted MRIs (Panels C). Hematoxylin-eosin staining showing accumulation of PCNSL cells within the perivascular space (Panels D and E). IHC of CD20 positive PCNSL cells (Panel F). Adapted from Ricard et al., 2012, von Baumgarten et al., 2018, and Kluin, et al., 2008.

Chapter 2

PCNSL cell biology, tumor microenvironment, and treatments

2.1 Tumor cells origin

2.1.1 The normal counterpart

The true nature of any cancer can only be evident in the light of its normal counterpart which, for most B-cell lymphomas including PCNSL and DLBCL, has been GC-experienced B-cells (Milpied et al., 2018; Steven H. Swerdlow et al., 2008; Steven H. Swerdlow et al., 2016). B-cells are key central elements that protect against an almost unlimited variety of pathogens thanks to their antibodies; unfortunately, defects in B-cell development, selection, and function can lead to malignancy. The B-cell lymphopoiesis is a multi-stage process with the expression of different transcription factors (TFs) and microenvironmental influences that starts in the bone marrow (BM) where hematopoietic stem cells (HSCs) differentiate into multipotent common lymphoid progenitors. Long-term HSCs have the ability to self-renew and reconstitute the entire immune system by differentiating into short-term HSCs which can then differentiate into multipotent progenitors (MPPs) that branch later into common myeloid progenitors and lymphoid-primed multipotent progenitors, having the latest one the potential to differentiate into natural killer (NK) cells, T or B lymphocytes (Barrios, Meler, & Parra, 2020; Pieper, Grimbacher, & Eibel, 2013). After commitment to the B-cell lineage, additional differentiation steps lead to the formation of pro-B and pre-B cells, which are the early B-cell precursors for immature and GC B-cells (Figure 2.1, I).

The pro-B cell stage is the initial paddle for the recombination of the non-contiguous germline variable (V), diversity (D), and joining (J) immunoglobulin (IG) gene segments; such process is refereed as V(D)J recombination. V(D)J rearrangements of the heavy chain (H-chain) together with those of the light chain (L-chain) generate a B-cell repertoire expressing antibodies capable of recognizing more than 5×10^{13} different antigens (Figure 2.2) (Menzel et al., 2014). The

D_H and J_H gene segment's rearrangements happen at pro-B which is followed by a joining upstream V_H change that leads to the early pre-B cell stage. During this stage and after 1-2 cell divisions, the pre-B-cell receptor (BCR) is formed by linking functional VDJ-constantμ pair with V-preB and λ-like, however, it is not yet detected on the surface. The pre-BCR has two roles, the first being to stop the H-chain rearrangements (allelic exclusion) and the second to initiate the rearrangement of the L-chain genes. The rearrangement of V_L and J_L segments allows replacing the V-preB/λ of the pre-BCR pair with the H-chain to form and express IgM on the cell surface of, now, immature B-cells. Immunoglobulin M generates drastic expression changes and initiates egress into the circulation system where immature B-cells will eventually reach the spleen. Within this secondary lymphoid organ, these now called transitional B-cells have increased IgM/D expression and receive survival signals through B-cell activating factor receptor (BAFF-R) which dictate their subsequent fate (naive, follicular, or marginal zone B-cells) (Barrios, Meler, & Parra, 2020; Pieper, Grimbacher, & Eibel, 2013). Interestingly, these transitional B-cells were found to express specific IgV_H regions, such as $V_{4\text{-}30\text{-}2}$, $V_{1\text{-}8}$, $V_{3\text{-}11}$, $V_{4\text{-}61}$, and $V_{4\text{-}31}$ (Wu et al., 2010) (Figure 2.1, I and II).

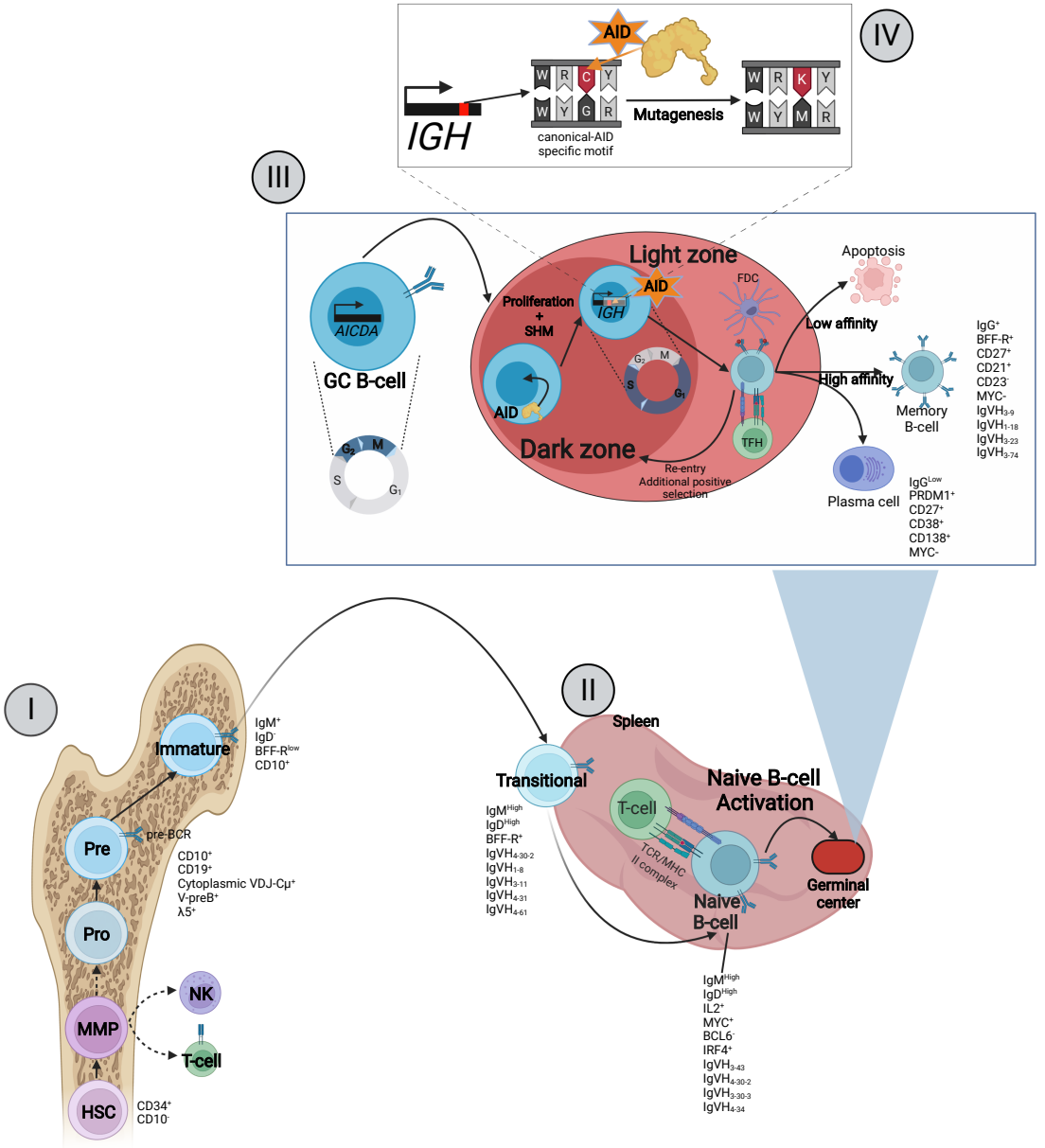


Figure 2.1: B-cell development and differentiation. B-cell commitment starts in the bone marrow, where hematopoietic stem cells differentiate into multipotent progenitors and common lymphoid progenitors, which then commit to the B-cell lineage and give rise to precursor B-cells (I). These precursors gradually rearrange their immunoglobulin genes and differentiate into mature naive B-cells, which leave the bone marrow to enter the bloodstream. Resting naive B-cells transit through lymph nodes and (II), eventually, they are activated by specific antigens via activation of the BCR, which induces the GC reaction. GC B-cells further rearrange and mutate their IG genes, rapidly proliferate and differentiate. Finally, the GC reaction gives rise to plasma cells producing large amounts of high-affinity antibodies and memory B cells (III). AID motifs of action are shown in IV.

Upon activation by a cognate antigen driven by T-helper (TH) cells, naive B-cells transiently express MYC due to the transcriptional inhibition of BCL6 (thanks to IRF4 and IL2) which ultimately leads to entering the germinal centers and to expressing activation-induced cytidine deaminase (AID) during the G2-M phases of the cell cycle. AID (encoded by *AICDA*) acts during the G1-S phases of the cell cycle and is responsible for somatic hypermutation (SHM), happening within the dark zone (DZ) of the GC, and class-switch recombination (CSR), happening within the light zone (LZ), having a role in further diversification of the variable or switch domains of IG genes (Endo et al., 2007; Muramatsu et al., 2000; Q. Wang et al., 2017). GC B-cells can undergo several rounds of affinity maturation by cycling through the DZ and LZ, with the help of follicular dendritic cells (FDC) and follicular helper T-cells (TFH), until high-affinity plasmablasts or memory B-cells are produced. On the contrary, low affinity or self-reactive cells fate apoptosis (Barrios, Meler, & Parra, 2020) (Figure 2.1, III). Finally, B-lymphocyte-induced maturation protein 1 (BLIMP1) suppresses MYC expression in plasmablasts and induces plasma cells differentiation. Regarding the use of IgV_H regions, $V_{3\text{-}43}$, $V_{3\text{-}30\text{-}3}$, $V_{4\text{-}30\text{-}2}$, and $V_{4\text{-}34}$ are frequently found in naive B-cells, meanwhile $V_{3\text{-}9}$, $V_{1\text{-}18}$, $V_{3\text{-}23}$, and $V_{3\text{-}74}$ in mature B-cells (Wu et al., 2010).

AID deamination of cytosine to uracil happens during the transcription of the IG genes and within specific deoxyribonucleic acid (DNA) motifs (Figure 2.1, IV) (Branton et al., 2020). When these lesions are resolved by the error-prone DNA polymerase-η, mutations can arise as A->C at WA motifs (W = A/T), which has been defined as non-canonical AID (Catalogue Of Somatic Mutations In Cancer (COSMIC) signature 9), or as C->T/G at WRCY motifs (R = purine; Y = pyrimidine) when resolved by base excision repair (BER) or mismatch repair (MMR) pathways, defined as canonical-AID (c-AID) throughout this thesis (Australian Pancreatic Cancer Genome Initiative et al., 2013; Delgado et al., 2020; PCAWG Mutational Signatures Working Group et al., 2020). The tight control of *AICDA* expression and AID activity is of primordial importance not only to ensure an antigenic fingerprint but to avoid AID off-target mutations, which have been addressed as blameworthy of lymphomagenesis (Chapuy et al., 2018; Fukumura et al., 2016).

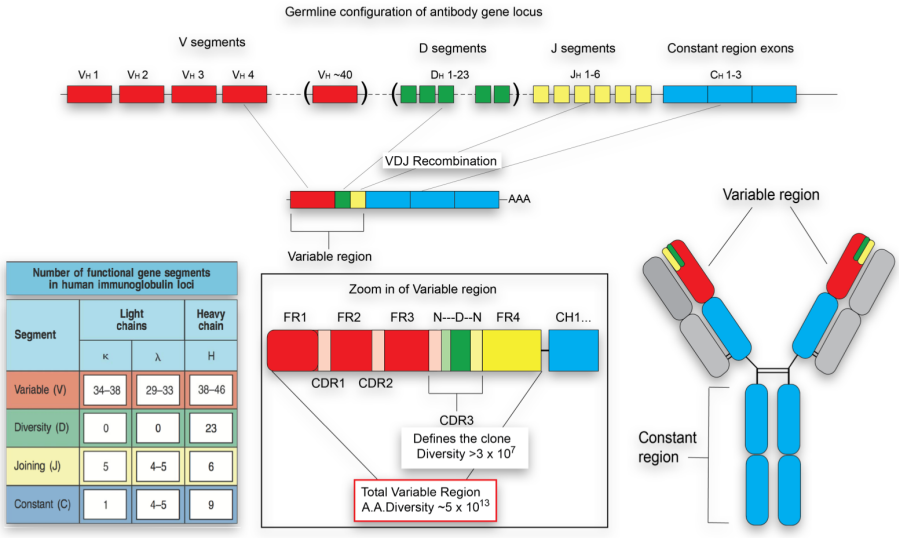


Figure 2.2: B-cell repertoire diversity. Germline configuration of antibody gene locus that can give a total diversity of $5x10^{13}$ different antigens. Adapted from Menzel et al., 2014.

From a molecular point of view, recent studies have demonstrated that the B-cell maturation process undergoes intensive gene methylation and expression changes to finally give rise to plasma cells and memory B cells that play an essential role in adaptive immunity. Kulis et. al. performed whole-genome bisulfite sequencing (WGBS) and high-density microarrays on ten subpopulations spanning the entire B-cell differentiation program and found that early differentiation stages mainly displayed enhancer demethylation, which was associated with up-regulation of key B-cell TFs (e.g. *ARID3A*, *BCL2*, *BLK*, *EBF1*, and *IRF4*) and affected multiple genes involved in B-cell biology. On the contrary, late differentiation stages have extensive demethylation of heterochromatin and methylation gain of polycomb-repressed areas and did not affect genes with apparent functional impact in B-cells. Interestingly, the authors also showed that the changes went into accumulative patterns in which each B-cell maturation stage, although characterized by a particular signature, kept an epigenetic memory of past differentiation stages (Kulis et al., 2015).

2.1.2 The abnormal counterpart

The correct identification of the B-cell maturation stage and of the location of the related cell type within the lymphoid follicle, from which a B-cell neoplasm derived, is the main principle behind the WHO classification of these tumors. Additionally, it is known that cancer cells, although victims of dramatic cellular identity alteration, maintain molecular imprints of the cellular lineage and the maturation stage from which they originated (Duran-Ferrer et al., 2020). For

example, B-cell acute lymphoblastic leukemia derives from pre B-cells, follicular lymphoma (FL) from the lymphoid follicle, marginal zone lymphoma (MZL) from the follicle's marginal zone, mantle cell lymphoma (MCL) from the mantle zone, DLBCL from the GC, Burkitt lymphoma (BL) from DZ B-cells, and multiple myeloma (MM) from terminally differentiated plasma cells (Figure 2.3) (Carbone et al., 2019). Initial understanding of B-cell malignancies assumed that they are "frozen" at a given B-cell differentiation stage arising in a particular location of the B-cell follicle, defined as cell of origin (COO) on the basis of classic histological definitions and gene expression profiling. Regarding PCNSL COO classification, the WHO termed it as a "late germinal centre exit B-cell arrested in terminal B-cell differentiation that shares genetic characteristics with both activated B-cells and germinal centre B-cells," previously defined in DLBCL as activated B-cell–like (ABC) or germinal center B-cell–like (GCB), respectively (Alizadeh et al., 2000; Steven H. Swerdlow et al., 2008; Steven H. Swerdlow et al., 2016). The GCB subtype corresponds to B-cells that are arrested at various stages of the GC transit (from dark zone to light zone B cells), whereas, the ABC to GC B-cells en route to plasma cell differentiation, resembling plasmablasts (Alizadeh et al., 2000; Carbone et al., 2019). Furthermore, as these definitions were found to be clinical predictors to response in DLBCL (GCB has better response than ABC), they are routinely used in the clinics (mainly for DLBCL since most PCNSL are ABC) with the help of either the Hans' algorithm (IHC profiling of CD10, BCL6, and MUM-1) or gene expression profiling (GEP) (Alizadeh et al., 2000; Camilleri-Broët et al., 1998; Hans, 2004).

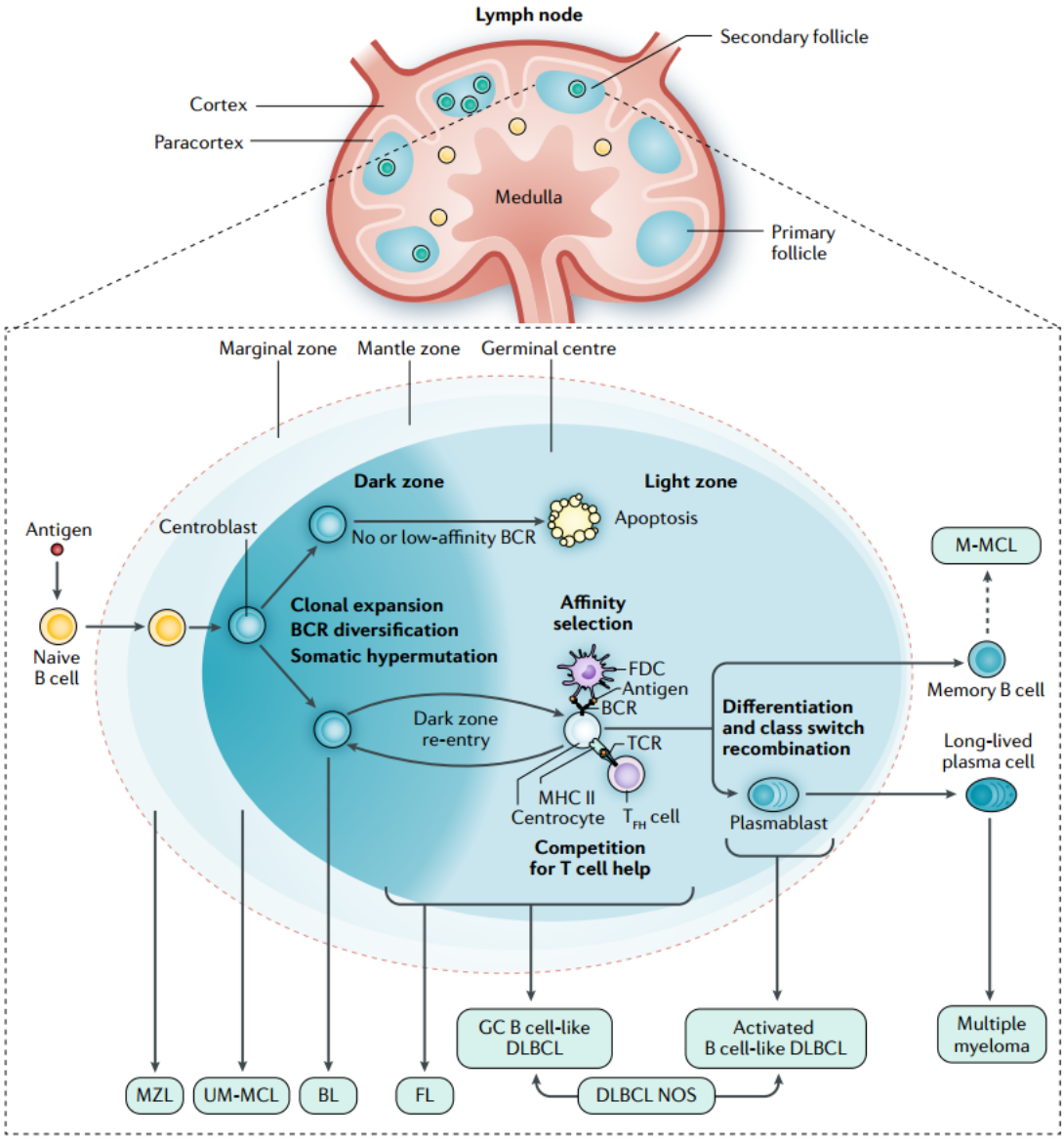


Figure 2.3: Normal-abnormal origin of mature B-cell lymphomas. B-cell neoplasms are the result of acquired malignancy at various stages of ontogeny. Most B-cell differentiation steps are associated with a malignant B cell subtype (defined as the cell of origin). The COO model assumes that FL is a follicle-related GC-derived B-cell, unmutated mantle cell lymphoma (UM-MCL) originates from mantle zone B cells, MZL resembles marginal zone B cells whereas Burkitt lymphoma (BL) resembles dark zone B cells. In the case of DLBCL, hence PCNSL, COO subtypes can either be ABC or GCB. TCR, T cell receptor. Taken from Carbone et al., 2019.

Thanks to next-generation sequencing (NGS) profiling of B-cell lymphomas, molecular precisions or improvements of these COO definitions have arisen. For example, it has been found that lymphomas "frozen" in the GC, that is FL, DLBCL, and PCNSL, present a higher number of mutations provoked by AID than

other types of lymphoma (Chapuy et al., 2018). *BCL2/BCL6* gene translocations with IGH regulatory regions lead to an ectopic overexpression of the protein, ultimately provoking iterative GC re-entries and hence clonal expansions, genomic instability, mutation acquisition (AID off-targets), and lymphomagenesis (Carbone et al., 2019; Chapuy et al., 2018). Furthermore, more recent studies using single-cell RNA-sequencing (scRNA-seq) of FL and DLBCL have found that malignant cells are not 'frozen' at a particular GC maturation stage but are rather 'dynamic,' as they showed functional diversity and consisted of multiple cell states that are co-existing within a single patient's tumor (Holmes et al., 2020; Milpied et al., 2018; Roider et al., 2020). Moreover, immunodeficiency-associated PCNSL have been recently found to have a more GCB-like phenotype instead of ABC (commonly found in immunocompetent PCNSL) in conjunction with a tolerogenic tumor microenvironment (TME) to adapt to an immunogenic virus (Gandhi et al., 2021). Additionally, Duran-Ferrer et. al. analyzed methylation data from normal and tumoral human cells from the B-cell lineage finding disease-specific hyper- and hypomethylation imprints. The authors then proposed a machine learning-based diagnostic algorithm named epigenetically-determined Cumulative MIToses (epiCMIT) which reflects the relative accumulation of mitotic cell divisions of a particular sample, including the mitotic history associated with normal cell development as well as with malignant transformation and progression (Duran-Ferrer et al., 2020).

2.2 Molecular definitions of PCNSL

Throughout the years, there has been accumulating progress in the understanding of the molecular parthenogenesis of PCNSL. In the light of NGS massive adoption, scientists have been able to gather more molecular data, such as genetic alterations (e.g. indels, mutations, amplifications, deletions, translocations), methylation, RNA expression (e.g. mRNA, miRNA, lncRNA), and protein expression, which enabled them coming up with molecular definitions of the disease. Three main molecular definitions of PCNSL, which are complementary as more data is obtained, have been proposed: I) The 2001 WHO classification that was later revised in 2017, II) Chapuy et. al. definition in 2018, and III) the 2020 definition by Wright et al (adapted from Schmitz 2018) (Chapuy et al., 2018; Schmitz et al., 2018; Steven H. Swerdlow et al., 2008; Steven H. Swerdlow et al., 2016; G. W. Wright et al., 2020).

The WHO classification defines PCNSL as B-cells blocked at terminal differentiation stages (see Chapter 2.1.2) with ongoing SHM that leads to aberrant somatic hypermutation (ASHM) (i.e. AID off-target's mutagenesis) of *IGH*, *PIM1*, *PAX5*, *TTF*, *MYC*, *KLH14*, *OSBPL10*, *SUSD2*, *BCL2*, and *BCL6* (master regulator of the GC reactions). Moreover, the finding of *BCL6* as recurrent translocation partner with the *IG* loci (17-47%) along with *PRDM1* point mutations (19%) would stop plasmacytic differentiation. PCNSL also presents fixed IgM/lgD phenotype in part due to miscarried IG class-switch rearrangements during which the

sμ region was deleted. PCNSL is characterized by constitutive Nuclear Factor-$\kappa\beta$ (NF-$\kappa\beta$) activity driven by alterations in genes of the BCR pathway (*CD79B* in 20%, *SHIP*-25%, *CBL*-4%, and *BLNK*-4%), of the toll-like receptor (TLR) pathway (*MYD88* L265P-50%) and others (*CARD11*-16%, *MALT1*-43%, *BCL2*-43%). The BCR complex, consisting of the IG H/L-chains as well as of CD79A and CD79B subunits, is indispensable for B-cell survival since it induces differentiation, proliferation, and apoptosis of B-cells. The BCR pathway also transmits its signals to the CARD11–BCL10–MALT1 (CBM) signalosome complex. Given its immune-privileged location, PCNSL escapes immune recognition by inducing loss of the 6p21 (37%) region harboring the MHC encoding genes. Finally, attributed genetic changes includes hypermethylation of *DAPK1* (84%), *CDKN2A* (75%), *MGMT* (52%), and *RFC* (30%) (Figure 2.4) (M. Deckert et al., 2011; Martina Deckert, Montesinos-Rongen, Brunn, & Siebert, 2014; King et al., 2020; Steven H. Swerdlow et al., 2008).

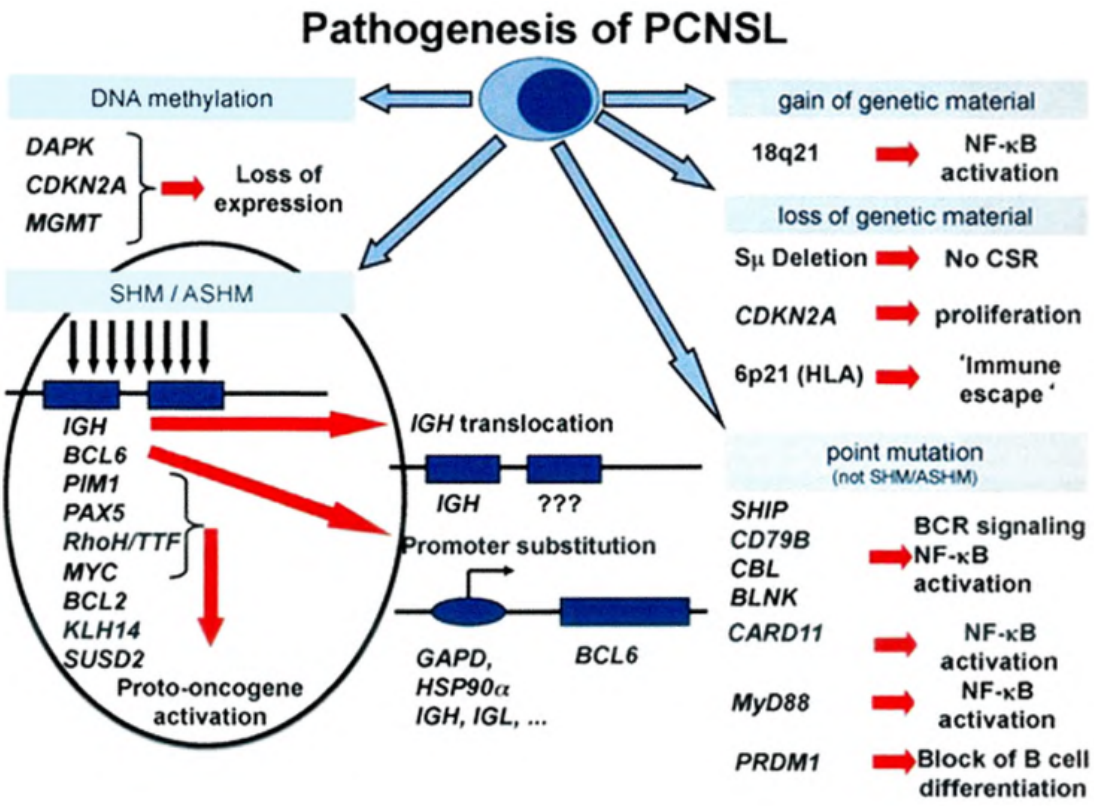


Figure 2.4: Pathogenesis of PCNSL. WHO recapitulation of PCNSL genetic alterations that ultimately lead to uncontrolled proliferation, impairment of apoptosis or B-cell differentiation, and immune escape. Taken from Kluin, et al., 2008.

The second definition was introduced in 2018 by Chapuy et. al. as result, principally, of one Blood (2016) and one Nature Medicine (2018) article (Chapuy et al., 2016, 2018). Their first publication, along with corroboration of previous molecular findings, described mutations in *IRF4* (29%), *ETV6* (21%), *BTG1*

(43%), and *TBL1WR1* (36%); copy gains of 3q12.3 (59%), 9p24.1 (52%) and 19q13.42 (32%); copy losses of 9p21.3 (71%); and chromosomal rearrangements of *ETV6* (13%) and *PD-L1/PD-L2* (13%). Interestingly, while *NFKBIZ* (3q12.3 copy gain) encodes a I$\kappa\beta$-z which coactivates canonical and noncanonical NF-$\kappa\beta$ pathways, *CDKN2A* (50% biallelic loss) encodes for p16 which is a cyclin D inhibitor necessary for the cell cycle arrest (Chapuy et al., 2016). In 2018, the same group carried out a comprehensive consensus clustering using recurrent mutations, copy-number alterations (CNAs), and structural variants (SVs), coming from 304 primary DLBCLs (FFPE tissue), that allowed them to identify five robust DLBCL subsets. They pinpointed cluster 5 (C5) to be systemic DLBCLs with CNS or testicular involvement since it showed characteristic molecular features of extranodal tropism (e.g. $MYD88^{L265P}$, *CD79B* clonal mutations). Moreover, they also noted C5 to have the highest contribution of c-AID activity, to practically be ABC COO subtype (96%), and to have the less favorable outcome compared to the others (Chapuy et al., 2018).

One caveat of the C5 definition is the lack of gene expression data which was later amended by Wright et. al. with their work "A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications." Using mutations, translocations, and CNA, the authors identified seven probabilistically defined DLBCL subtypes, to which gene expression, B-cell differentiation, IgV_H expression, and TME profiles, were later added. PCNSL belongs to the so-called "MCD" (based on the co-occurrence of $MYD88^{L265P}$ and *CD79B* mutations) genetic subtype whose added molecular features are: TFs overexpression (*IRF4*, *OCT2*), increased oncogenic signatures activity (proliferation, NF-$\kappa\beta$, JAK1 kinase, and MYC induction), heterogeneous B-cell differentiation signatures (DZ, LZ, and intermediate zone GC programs), increased IgM expression, self-antigen-dependent chronic active BCR signaling (high $IgV_{H4\text{-}34}$ expression), and a cold TME (see Chapter 2.3) (G. W. Wright et al., 2020). The most important contributions of each definition are summarized in Table 2.1, while the most extensively studied signaling pathways alterations are illustrated in Figure 2.5.

Table 2.1: Summary of PCNSL molecular definitions.

Feature[a]	WHO (2017)	Chapuy (2018)-C5 cluster	Wright (2020)-MCD subtype
Genetic themes	NF-kB activation	NF-kB activation mediated by NFKBIZ	
	BCR signaling disruption		
	CDKN2A mediated proliferation	Proliferation by CDKN2A biallelic loss	
	CBM complex alterations	Immune escape by PD-L1/PD-L2 translocations	
	6p21 mediated immune escape	High c-AID activity	
	Disfunctional CSR		
	BCL6 translocations		
	ASHM		
Epigenetic themes	DAPK, MGMT, and CDKN2A hypermethylation	—	—
B-cell differentiation stage	'Frozen' at GC	GC mostly ABC subtype	DZ, LZ and intermediate zone GC
TME	—	—	Cold
Gene expression signatures	—	—	High IgM expression
			Proliferation
			NF-kB
			MYC induction
			Autoreactive BCRs

[a]The described features are accumulative, meaning that the molecular themes of an author's molecular definition are part of the next molecular definition as these features were corroborated each time.

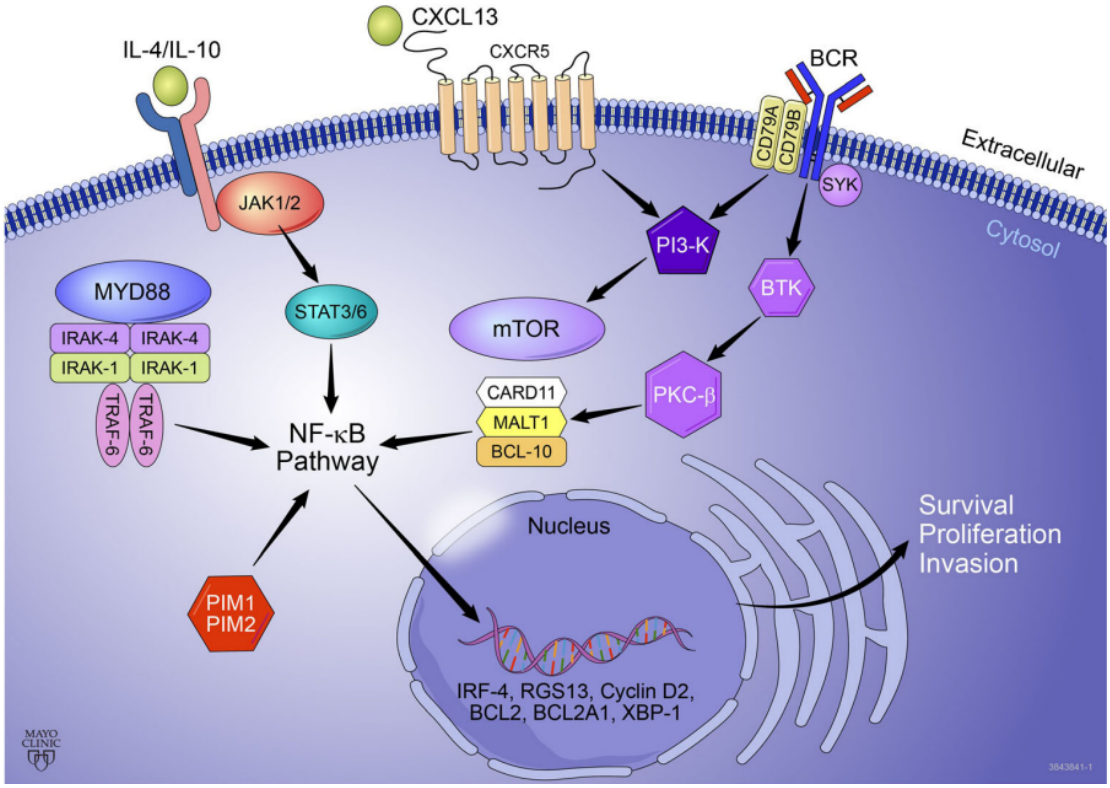


Figure 2.5: Main signaling pathways disrupted in PCNSL. NF-$\kappa\beta$ signaling can be activated via MYD88, JAK/STAT (via IL-10/IL-4 increased levels), CBM complex, and/or BCR pathway alterations. Altogether, these aberrations lead to increased pro-survival, proliferation, and invasion. Taken from King, et al., 2020.

In spite of the molecular insights these definitions have brought, PCNSL heterogeneity has not been properly addressed mainly due to the lack of a large number of patients and multi-omic data integration, i.e. having distinct types of molecular information (e.g. methylation, mutations, CNAs, gene expression, tumor location, TME, etc) for the same cohort. The WHO classification compiles the findings of different scientists on the field, but unfortunately, their cohort's size did not exceed 40 patients, had only one-omic level of information and some lacked clinical information. Additional molecular features were included with the introduction of the C5 and MCD molecular definitions, however, the studies were done in the context of DLBCLs to better stratify DLBCLs and not to address the PCNSL heterogeneity since their studies pulled together all types of extranodal lymphomas. Moreover, due to the difficulty of acquiring FF tissue, most of the studies have been performed after genetic material obtention from FFPE tissue which is subject to chemical degradation.

2.3 The tumor microenvironment

The TME plays an ineluctable and remarkable role in tumor biology and is defined as a cellular (i.e. blood vessels, immune cells, and fibroblasts), molecular (i.e. intercellular signaling molecules, extracellular matrix (ECM)), and dynamic network surrounding tumor cells. Substantial accumulating proof suggests that tumor development is not only due to the accumulation of intrinsic abnormalities but also to extrinsic signals from the TME. Both the tumor and the TME are engaged in a continuous, interactive, and dynamic cross-talk via various signaling pathways (Broekman et al., 2018). PCNSL represents a special scenario as the immune reactions to the malignant B-cells happen within an "immuno-privileged" organ (i.e. region less subject to triggering immune responses), where the presence of the blood-brain barrier (BBB) greatly limits the exchanges, of cells and molecules, between the brain and the blood vessels (Martina Deckert, Montesinos-Rongen, Brunn, & Siebert, 2014). The present chapter aims to give a general overview of the antitumor immune response, describe the tumor escape mechanisms in general and within lymphomas, and recapitulate the most relevant findings of the TME in DLBCL and PCNSL.

2.3.1 The antitumor immune response

The accumulation of genetic alterations giving rise to cancer cells can also trigger the integrated response between the innate and adaptive. Proteasome-mediated degradation of cancer proteins generates 8- to 12-mer peptides which can be bound to MHC-I molecules on the surface of cancer cells, so-called neoantigens (Schmidt et al., 2021). These neoantigens are distinguishable from their normal counterparts as they are the result of encoding mutated genes that initially gave a biological advantage to the tumor cells. It is known that these cancer-specific peptide-MHC-I complexes can be recognized by $CD8^+$ T-cells to initiate an immune response, however, they rarely provided protective immunity nor could they be mobilized to provide a basis for therapy (D. S. Chen & Mellman, 2013). To produce an effective antitumor response (endogenously or therapeutically) a series of stepwise events must take place (Figure 2.6): the produced neoantigens are detected and captured by dendritic cells (DCs) for later presentation (step 1). Of note, pro-inflammatory molecules and chemokines released by the tumor cells themselves will recruit innate immune cells to this local source of "danger." DCs present the neoantigens on MHC molecules to T-cells (step 2), triggering the activation and the priming of effector T-lymphocytes against tumor-specific antigens (step 3). The ratio of T-effector lymphocytes to T-regulatory lymphocytes presents a critical determinant in the response. Finally, these cytotoxic T-lymphocytes (CTLs) exit lymph nodes and travel through the bloodstream (step 4) to infiltrate the tumor bed (step 5). CTLs recognize antigenic peptide-MHC complexes by T-cell receptor (TCR) interactions (step 6) and proceed to cancer cells destroying by means of releasing perforin and granzyme (step 7). The death of cancer cells releases additional tumor antigens that re-initiates the cycle to amplify the T-cell response (D. S. Chen & Mellman,

2013; Y.-T. Liu & Sun, 2021).

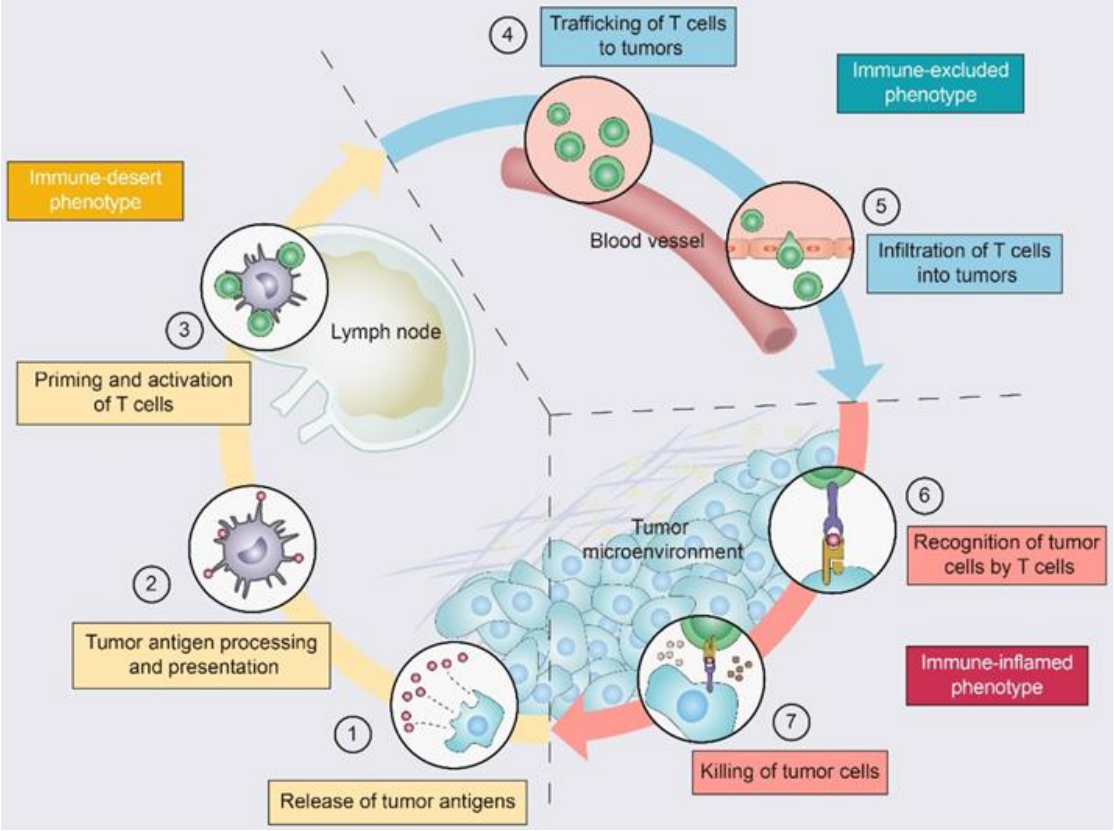


Figure 2.6: Cancer immunity cycle and cancer immunephenotypes. Antitumor immunity is mediated to a large extent by CTLs and can be divided into seven major steps. Tumors with the immune-desert phenotype (yellow) cannot pass steps 1-3 due to the absence of CTLs in both the tumor and its margins. Tumors with the immune-excluded phenotype (blue) cannot exceed steps 4-5 due to a lack of CTLs in the tumor bed. Tumors with the immune-inflamed phenotype (red) cannot exceed steps 6-7 due to T-cell exhaustion and checkpoint activation. Taken from Liu and Son, 2021.

Furthermore, according to the spatial distribution of CTLs within the tumor and the TME, tumors can present one of three immunophenotypes: immune-inflamed, immune-excluded and immune-desert. "Hot" or "immune-inflamed" tumors present high T-cell infiltration, increased interferon-γ (IFN-γ) signaling, expression of PD-L1 and high tumor mutational burden (TMB). Most commonly known as "Cold" tumors, immune-excluded tumors and immune-desert tumors, are characterized by low TMB, low MHC/PD-L1 expression, and presence of immunosuppressive cell populations such as tumor-associated macrophages (TAMs), T-regulatory cells (Tregs), and myeloid-derived suppressor cells (MDSCs). The difference between immune-excluded and immune-desert tumors is that CTLs are localized at invasion margins, for the first one, but absent in the latter (Y.-T. Liu

& Sun, 2021) (Figure 2.7).

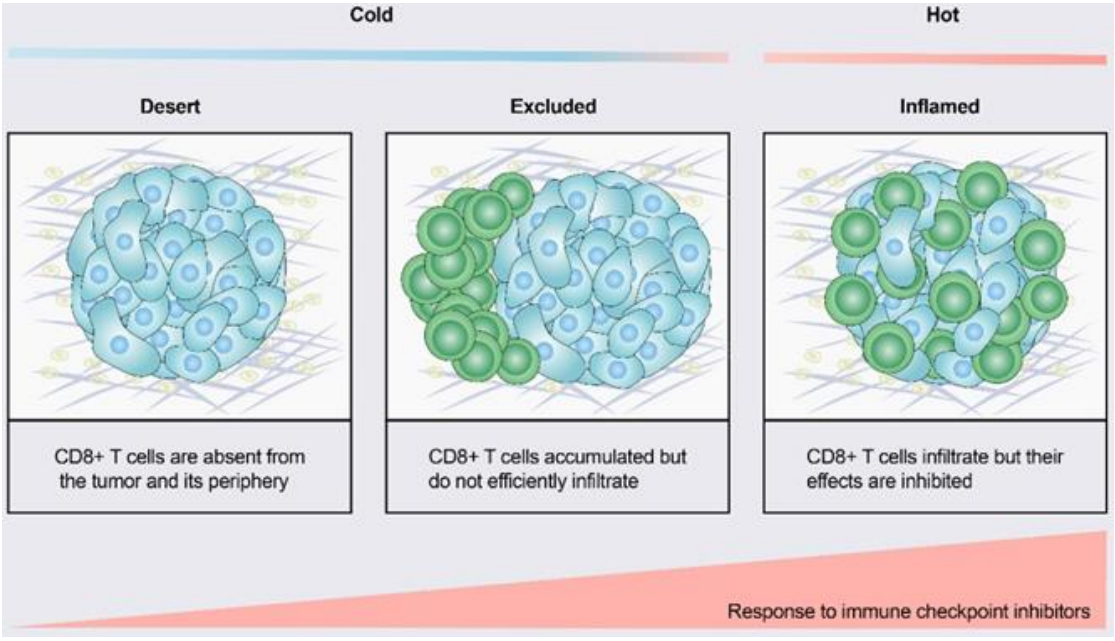


Figure 2.7: Characteristics of cancer immunephenotypes. Spatial distribution of CTLs within the TME distinguishes three immunephenotypes with different response rates to immune checkpoint inhibitors. Taken from Liu and Son, 2021.

Depending on the specific TME and immunophenotype (cold or hot), tumors interrupt or slow down the cancer immunity cycle through different tumor escape mechanisms (Y.-T. Liu & Sun, 2021).

2.3.2 Tumor escape mechanisms

The breach of the host's immune defenses is likely provoked by genetic/epigenetic changes induced by the tumor cells that confer resistance to detection and elimination by the immune system. Tumor cells possess a menu of choices to escape from the immune system that can be generally categorized according to the cancer immunity cycle step they are bypassing in (Figure 2.6). The objective of this section is to present the escape mechanisms used by cold/hot tumors that have actually been found to be used by DLBCL or PCNSL tumors.

Cold tumor mechanisms

As cold tumors lack T-cell infiltration, there are many factors that can influence T-cell priming and T-cell homing to the tumor bed (steps 1-5 of the cancer immunity cycle), leading to a noninflamed T-cell phenotype and failed antitumor immunity.

Antigens production

The release of tumor antigens (step 1) can be disrupted in terms of quantity and quality (affects mostly step 3 and 6). Antigens can either be nonmutated self-antigens or neoantigens generated by nonsynonymous somatic mutations. Self-antigens come from overexpressed nonmutated proteins that are commonly undetected by the immune system given that they could be also produced by their normal counterpart. On the other hand, neoantigens are more tumor-specific and hence may promote T-cell priming and infiltration which can lead to a response. It has been long recognized that high TMB (i.e. total nonsynonymous single-nucleotide mutations in coding regions) leads to higher neoantigen load which makes the tumor more likely to prime the immune system. Furthermore, this high neoantigen burden has been positively correlated with CTL infiltration in different tumor types including DLBCL (Fangazio et al., 2021; Rooney, Shukla, Wu, Getz, & Hacohen, 2015). Choosing inducing methylation or gene expression changes while maintaining a low TMB is one option in the menu to reduce, but not eliminate, immune recognition.

Another aspect to consider is the quality of the neoantigen produced, which can be affected by the clonality of the mutation and the sequence of the peptide. Regarding clonality, a mutation can be seen as clonal or subclonal depending on the temporal acquisition of the mutation, that is, during the early cancer evolution or at late times of cancer evolution. Clonal mutations, hence clonal neoantigens, are present in >= 95% of the total cancer cell population; while subclonal mutations happen in a minority of cells (subpopulations) (Shinde et al., 2018). These concepts are important given that immune recognition of clonal neoantigens is preferred by the immune system as it would be able to target and kill a broader range of cancer cells. The sequence of the neoantigen affects the affinity to both the MHC molecule on the surface of the antigen presenting cell (APC) or the cancer cell and the TCR on the surface of $CD4^+$ or $CD8^+$ T-cells (Schmidt et al., 2021).

Antigen presentation and priming of T-cells

MHC class I and class II molecules on APC present peptides at the cell surface to $CD8^+$ and $CD4^+$ T cells, respectively. Because $CD8^+$ T-cells are mostly associated with cancer cells' destruction and response, we will only focus on them. Forming a stable peptide-HLA-I (pHLA) complex depends on the neoantigen affinity to the HLA-I molecules which are encoded by highly polymorphic genes (HLA-A/B/C). Binding affinity and stability of the pHLA complex are affected by signals from cleavage and antigen transport, hotspots' presentation, gene expression of the source protein, clonality of the mutations for cancer neo-epitopes, and affinity of the non-mutated counterpart (competence) (Neefjes, Jongsma, Paul, & Bakke, 2011). After successful pHLA formation, the TCR needs to engage a stable immune synapse with the APC in which the peptide sequence also plays a role. A recent bioinformatic-experimental study using immunogenic and non-immunogenic peptides, experimental testing, and X-ray structures showed that TCR binding and recognition improves with the presence of hydrophobic amino acids (aromatic W, F, Y followed by V, L, and I) at specific "MIA" positions (position P4-PΩ-1) due to increased structural avidity, stacking interactions, hydrogen bond acceptance

and limited rotational freedom with the TCR (Schmidt et al., 2021).

Another escape mechanism of cancer cells is immunoediting which is the negative selection of mutations giving rise to immunogenic neoantigens. The main idea behind immunoediting is that cancer cells "know" the host's HLA repertoire, so they avoid giving rise to mutations that could, after the complex antigen processing, give rise to peptide sequences with higher probabilities of immune detection (Schmidt et al., 2021). In addition, alterations in the antigen processing and presentation machinery (APM) such as downregulation of MHC-I molecule expression, and absence of beta-2-microglobulin (B2M) or transporter associated with antigen processing (TAP) proteins can affect the antigen presentation process which further affects the priming of T lymphocytes. Tumor cells can affect the expression of these proteins by secreting products like NBR1, and IL-10 (Yamamoto et al., 2020).

The TME immunosuppressive cells and factors

T-cell priming and infiltration can also be disrupted by the presence of dense stroma and immunosuppressive cells and factors. Cancer-associated fibroblasts (CAFs) are predominantly located at the infiltrating edges of tumors, regulating tumor metastasis and influencing angiogenesis by synthesizing and remodeling the ECM and producing cytokines and transforming tumor margins into immune "cold" zones. Firstly, CAFs can produce TGFβ which limits the proliferation of $CD4^+$ T lymphocytes, induces its conversion into Tregs, and negatively affects DC differentiation and antigen-presenting functions (Brabletz et al., 1993; Shimabukuro-Vornhagen et al., 2012). Secondly, they can impede T-cells infiltration via ECM formation and CXCL12 production (Feig et al., 2013). At the same time, TAMs can affect T-cell recruitment by promoting abnormal angiogenesis through the production of VEGF and matrix metalloproteinase-9 (MMP9) (Lin & Pollard, 2007). Myeloid cell differentiation towards an immunosuppressive M2 macrophage phenotype can be achieved by either TAMs' secretion of cytokine colony-stimulating factor-1 or B-cells' secretion of GABA (Xia et al., 2020; Baihao Zhang et al., 2021). In the DLBCL context, CAFs were found to be part of a mesenchymal-like TME along with vascular endothelial cells (VECs) and fibroblastic reticullar cells (FRCs). In this context, CAFs produced a restrictive ECM composed mainly of collagens and proteoglycans which influenced DLBCL growth (Kotlov et al., 2021).

Hot tumor mechanisms

Already granted of an immune-inflamed phenotype, hot tumors escape options are bypassing steps 6 and/or 7 through inducing defects in tumor APM, reducing neoantigens quality for TCR engaging disruption, or promoting T-cell exhaustion.

Defects in recognition of cancer cells by T-cells

Immunoediting does not only serve to avoid antigen presentation and priming of T-cells by APC but also to avoid cancer cell recognition in later stages. Nev-

ertheless, on top of this, the most frequently described and opted mechanism in DLBCL and PCNSL is the induction of mutations or deletions in the APM. Around 50% and ~37% of DBLCL and PCNSL, respectively, present aberrant MHC-I expression mainly induced by monoallelic HLA disruptions (43% and 37%), biallelic HLA disruptions (9.4%, 1%), B2M focal deletions (14% and 20%), B2M biallelic disruption (26%, 0%) and/or TAP1/2 monoallelic loss (21% and 5%) (Chapuy et al., 2016; Fangazio et al., 2021). Moreover, the HLA loss has been proven to be haplotype-specific as the loss attends to eliminate the HLA haplotypes bearing the highest affinity for relevant tumor neoantigens' recognition. The lower biallelic loss of the MHC-I in PCNSL could be a defense mechanism to avoid NK elimination (Fangazio et al., 2021). Another recent study in DLBCLs demonstrated that MHC-I expression, negatively regulated by IRF4 and the PRC2 complex (transcriptional repressor involved in methylating the lysine in the position 27 (K27) of histone H3), can be restored with thymidylate synthase and EZH2 inhibitors to enhance tumor peptide presentation (Dersh et al., 2021).

T-cell exhaustion

After proper cancer recognition by TCR engaging to the pHLA complex, cancer cells can still escape by means of T-cell/NK exhaustion which is the loss of T-cell/NK effector function, including proliferation, the release of cytokines, and secretion of cytolytic molecules, due to continuous antigen stimulation (Saleh et al., 2020). Besides TME immunosuppressive cells and factors, T-cell/NK exhaustion can be induced by the overexpression of multiple immune checkpoint (IC) molecules such as PD-1, TIM-3, CTLA-4, GITR, TIGIT, ICOS, LAG3, CD28, CD27, VISTA, (inhibitory receptors on T-cells), CD96, CD94, TIGIT, LAG3 (on NK cells), and/or PD-L1, PD-L2, CD155, CD111, CD112, CD40L, ICOSL, CD80, CD86, and CD70 (associated ligands on APC or cancer cells, Figure 2.8) (Pardoll, 2012; Wherry & Kurachi, 2015).

Importantly, while T cell exhaustion prevents optimal control of infections and tumors, targeting these IC by immune checkpoint inhibitors (ICIs) can reverse this dysfunctional state and reinvigorate immune responses. While the use of ICIs (specially anti-CTLA-4, anti-PD-1, and anti-PD-L1) have revolutionized cancer therapy, heterogeneous responses are still observed across different cancer types, including PNCSL (see Chapter 2.4) (Garcilazo-Reyes et al., 2020; Pardoll, 2012). Today, because several previous studies found high TMB to be correlated with ICI response, the USA Food and Drug Administration (FDA) have approved it as a predictive biomarker. Moreover, a recent 2021 bioinformatic study combined with scRNA-seq analysis found clonal TMB along with CXCL9/CXCL13 expression to be the strongest predictors of ICI response. CXCL13 was observed to be expressed in clonal neoantigen-reactive CD8+ tumor-infiltrating lymphocytes (TILs) (Litchfield et al., 2021).

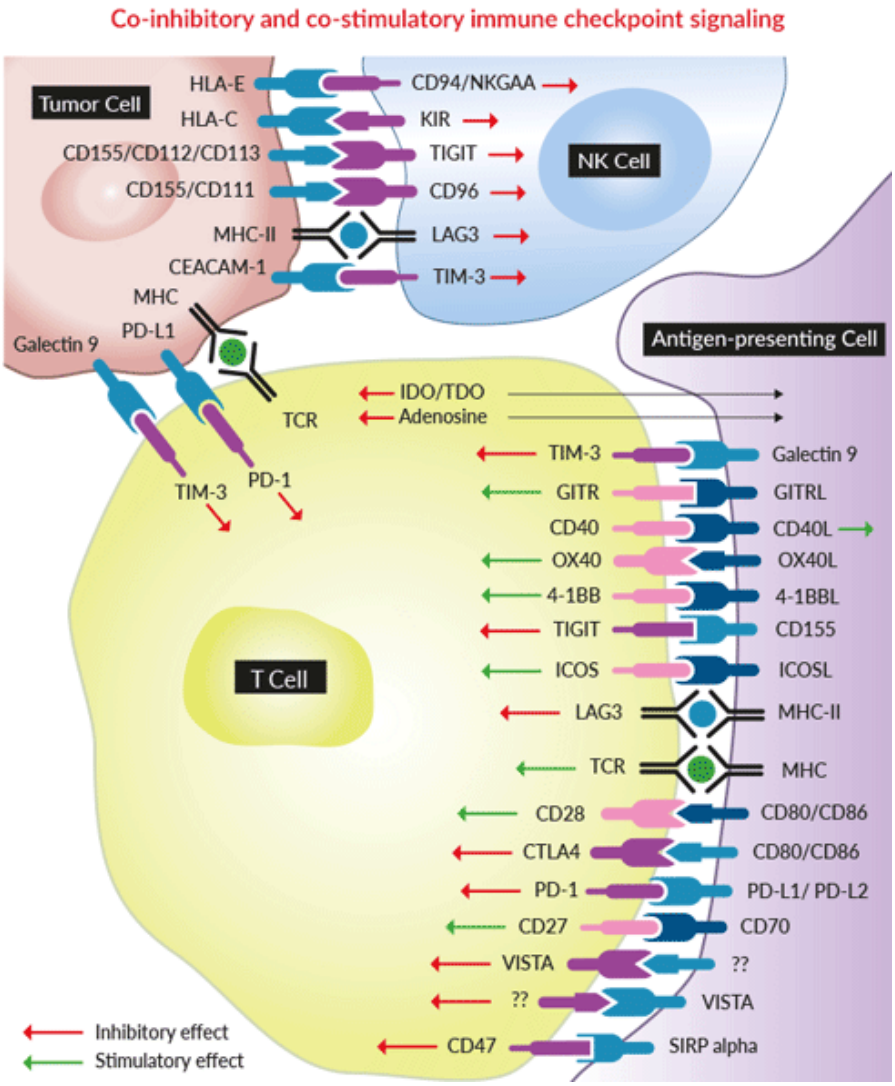


Figure 2.8: Immune checkpoint receptors and ligands associated with T/NK cells exhaustion. Co-inhibitory and co-stimulatory immune checkpoint signaling within T, NK, cancer and antigen-presenting cells. Taken from InvivoGen at www.invivogen.com/immune-checkpoints.

2.3.3 The TME in DLBCLs

Due to its biological similarity with PCNSL, it results incumbent to describe current knowledge about the DLBCL TME. Data from patients with lymphoma and animal models indicate that in the lymphoma niche, external stimuli provided by microenvironmental cells and the ECM contribute to disease development, progression, and response to treatment. However, the majority of the molecular and therapeutic studies of this disease have been focused on the characterization of the DLBCL cell as an isolated entity. By using genomic data from 4,655 DLBCLs, an early 2021 study revealed the existence of four different TME that exhibited distinct clinical behaviors. The found subtypes, termed GC-like, mesenchymal,

inflammatory, and depleted, have many hot-cold tumor characteristics and share some inter-tumoral genetic alterations per group (Figure 2.9A).

The GC-like subtype was characterized by lymphatic endothelial cells (LECs), FDCs, TFHs, and Tregs in the TME, along with *TNFRSF1*, *CD83*, *STAT6*, and *HSF1* genetic alterations. The mesenchymal subtype had a higher presence of VECs, CAFs, FRCs, and macrophages M1 (pro-inflammatory) in the TME; in addition, they also presented mutations in *E2H2*, *B2M*, *GNA13*, *GNAI2*, and *P2RY8*. The depleted subtype had genomic alterations leading to decreased p53 activity, perturbation of cell-cycle regulation (e.g. *CDKN2A* deletions), and high proliferative activity. Finally, the hot-like tumor (inflammatory subtype) was enriched in neutrophils, TAMs, macrophages M1, Tregs, TFHs, $CD8^+$ T-cells with high PD-1 expression (exhausted), and also NK, MHC-I, ICs, NF-$\kappa\beta$, JAK/STAT, and TNF activities (Kotlov et al., 2021).

Regarding clinical outcome at univariate level, Kaplan-Meier (KM) survival models using OS and progression-free survival (PFS) showed improved prognosis in GC-like followed by mesenchymal, inflammatory, and depleted subtypes (Figures 2.9B and C). The impact of TME remained as an independent prognostic factor even after adjustment by multivariate correction using the international prognostic index (IPI) (includes age and Karnofsky score) and the COO (Kotlov et al., 2021).

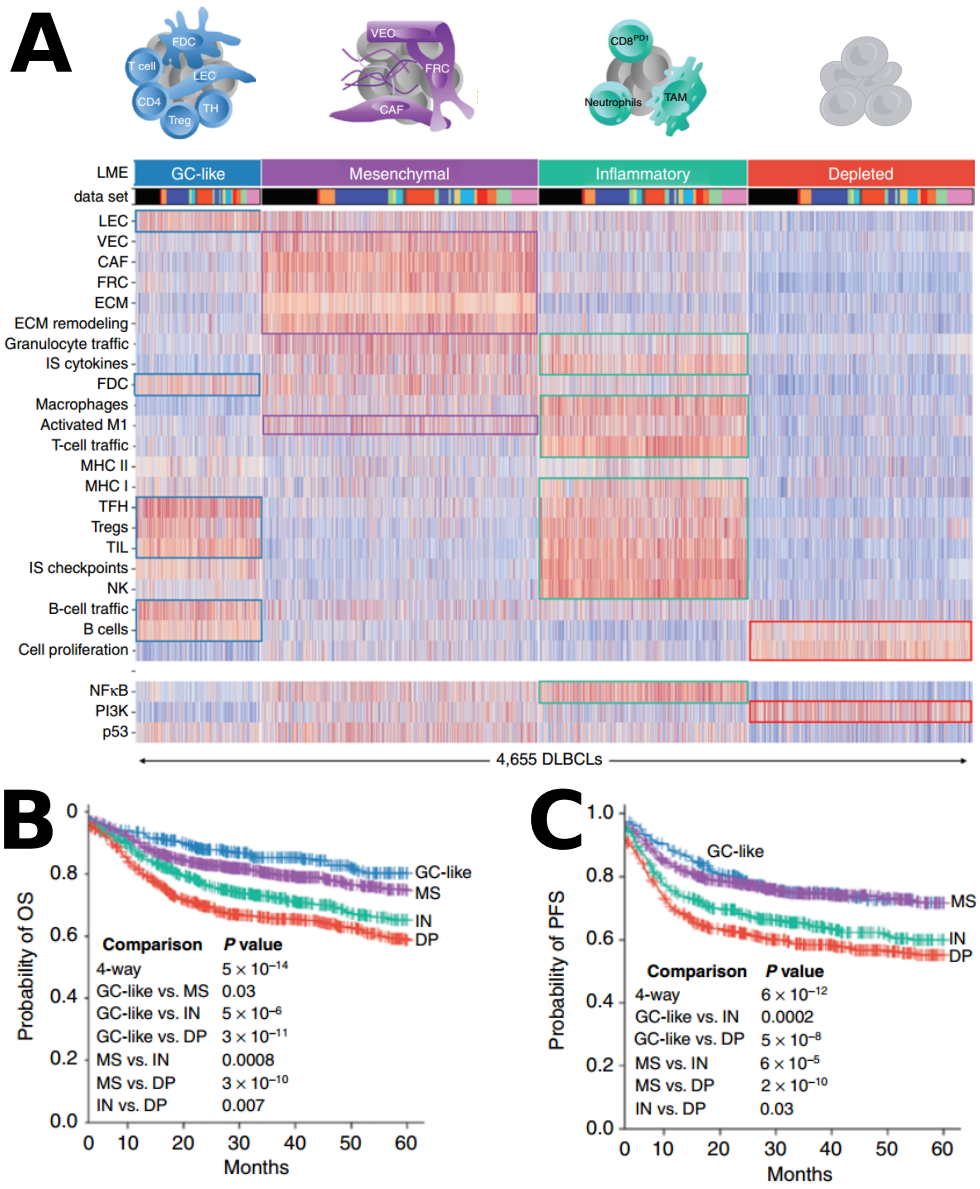


Figure 2.9: DLBCL TME subtypes have a distinct clinical impact. Heatmap of the transcriptional activity of distinct gene signatures denoting four major TME clusters termed GC-like, mesenchymal, inflammatory, and depleted (Panel A). Chemoimmunotherapy response in a balanced cohort (n = 105) with responsive and nonresponsive (refractory and relapsed) DLBCL using OS (Panel B) or PFS (Panel C). Adapted from Kotlov et. al., 2021.

2.3.4 The TME in PCNSLs

Even though the TME of certain brain/lymphoid tumors is starting to be unraveled, little is known about the cellular and molecular immune composition in PCNSL. The precise location of a PCNSL tumor likely drives the TME composition since it can develop in the brain parenchyma, but also the perivascular and

meningeal spaces (Figure 2.10). PCNSL is considered a tumor confined within an "immuno-privileged" zone because of the presence of the BBB. Nevertheless, recent studies have described a network of lymphatic vessels running parallel to the dural venous sinuses which allow for the drainage of cells and CSF into deep cervical lymph nodes. These CNS lymphatic vessels express all of the molecular hallmarks of LECs and carry around 24% of all sinusal T cells and 12% of all sinusal MHC-II$^+$ cells (Louveau et al., 2015). Additionally, the glymphatic system clears and carries the interstitial fluid (contains solutes and antigens) into the CSF to potentially elicit anti-tumor responses. In the steady-state, the immune system of the CNS is composed mainly of different types of macrophages such as microglia, meningeal, perivascular, and choroid plexus macrophages. Moreover, direct vascular connections between the meninges and the skull bone marrow have recently been demonstrated to serve as a private reservoir of myeloid cells and B-cells in the event of homeostasis and CNS injury (Alcantara, Fuentealba, & Soussain, 2021).

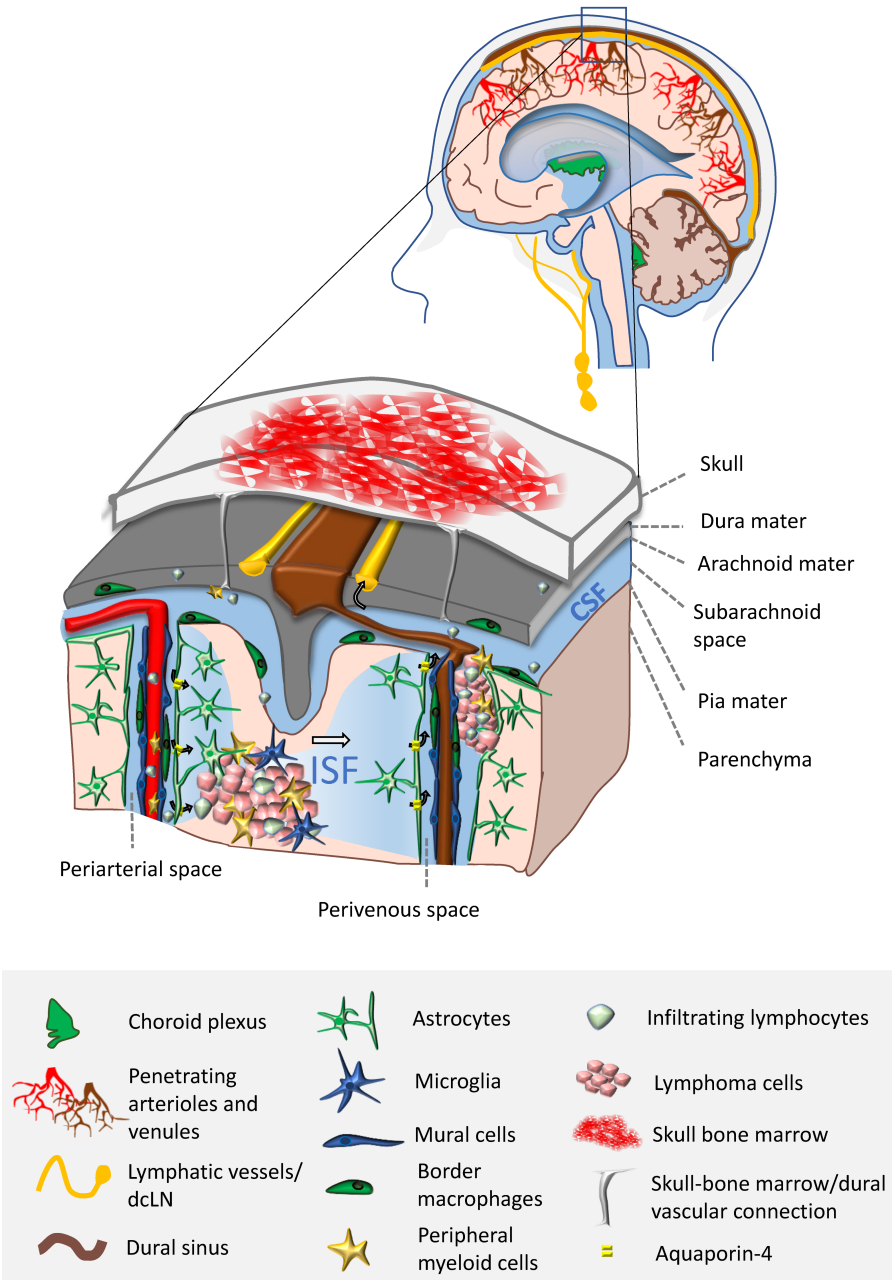


Figure 2.10: PCNSL TME configuration. There are three potential sources of immune cells within PCNSL TME: derived from resident populations, from the blood, and also from skull bone marrow reservoirs. PCNSLs developing within the CSF compartments (perivascular and meningeal spaces) interact directly with border macrophages, lymphocytes, the glia limitans (formed by astrocytic endfeet), endothelial cells, and mural cells (pericytes and smooth muscle cells). Inside the CNS parenchyma, tumor cells are in close contact with microglia, astrocytes, and infiltrating immune cells: lymphocytes and peripheral myeloid cells. Arrows indicate the directionality of CSF/Interstitial fluid bulk flow, which is facilitated by Aquaporin-4 expressed on astrocytes. Adapted from Alcantara et. al., 2020.

Several retrospective studies using either classical or high-plex IHC have pointed out correlations between the TME and outcome (Table 2.3). While initial studies have found a link between improved survival with perivascular T-cells or TILs, recent ones (using RNA-seq data) have found these TILs to be expressing IC receptors like PD-1 and TIM-3 (Alame et al., 2021; Four et al., 2017; He et al., 2013; Komohara et al., 2011; Kumari, Krishnani, Rawat, Agarwal, & Lal, 2009; Marcelis et al., 2020; Ponzoni et al., 2007). Interestingly, even though previous studies have found 9p24.1 amplification (involves PD-L1) recurrent, Marcelis et. al. did not find such amplification by means of FISH methodology; this demonstrates the need for a bigger PCNSL cohort. Furthermore, a globally increased ratio of M1/M2-like TAMs has been associated with a better outcome using either IHC or RNA-assisted immune deconvolution (H. Cho et al., 2017; Marcelis et al., 2020; Miyasato et al., 2018; Sasayama et al., 2016).

On the other hand, a recent transcriptomic analysis combining RNA-seq (n = 20) and microarrays (n = 34) described three immunophenotypes termed as rich, intermediate, and poor with OS implications. While the immune-rich group had the best OS and was characterized by a high number of $CD4^+$, $CD8^+$, Tregs, TAMs, and DCs, the immune-poor group was practically a cold-like tumor, and the intermediate group presented immune cells heterogeneity. Besides the TME, the authors found high STAT3, IFN-γ, TNF-α, MHC-I, and PD-L1 activity in the hot-like tumor; whereas, WNT/β-catenin, HIPPO, and NOTCH activity in the cold-like tumors (Alame et al., 2021).

The accumulative information provided by these studies has improved our understanding of the origin and characteristics of the PCNSL TME, however, the link between the intrinsic causative biologic factors of the disease (e.g. genetic/epigenetic alterations) is still missing.

2.4 PCNSL treatments and prognosis

The present section intends on describing the prognosis and current therapeutic strategies of PCNSL in light of the understanding of the previously described intrinsic and extrinsic factors encircling the biology of PCNSL. Firstly, this section describes the two major scoring systems widely used for PCNSL and compiles the prognostic markers associated with the disease. The second part briefly describes the current therapeutic strategy which can be divided into induction, consolidation, and maintenance phases. These phases primarily rely on high-dose (HD) methrotexate (MTX)-based regiment followed by either whole-brain radiotherapy (WBRT) or autologous hematopoietic stem cell transplantation (ASCT); however, the current understanding of PCNSL genetic alterations and its TME has allowed the inclusion of additional emerging maintenance/salvage therapies such as targeted therapies (e.g. ibrutinib, temsirolimus, buparlisib, thalidomide, temozolomide, and CAR-T therapy), TME-modulating therapies (e.g. ICIs, and immunomodulators), and BBB-permeabilizing therapies (Figure 2.11) (Hoang-Xuan et al., 2015; Schaff & Grommes, 2021; Steven H. Swerdlow et al., 2008; Steven H.

Table 2.3: Overview of published studies investigating the TME in PCNSL. Adapted from Marcelis et. al. 2020.

Study (Year)	Patients	Conclusions
Ponzoni et. al. (2007)	96	Presence of reactive perivascular CD3+ T-cell infltrate is associated with ↑ OS. Presence of tumor necrosis has no prognostic signifcance.
Kumari et al. (2009)	30	Reactive perivascular CD3+ T-cell infltrate shows no correlation with the diferent IELSG risk score groups. Tumor necrosis shows no correlation with the diferent IELSG risk score groups.
Komohara et al. (2011)	43	The number of infltrating CD68+, CD163+ and CD204+ TAMs has no prognostic signifcance.
He et al. (2013)	62	Presence of reactive perivascular CD3+T cell infltrate is associated with ↑ OS. Presence of aggregative perivascular tumor cells, stained with XBP1 and CD44, is associated with ↓ OS.
Chang et al. (2015)	62	PCNSL shows ↓ HLA-DR expression, ↓number of S100+dendritic cells, ↓ number of CD45RO+ efector or memory T-cells in comparison with DLBCL. ↓ number of infltrating granzyme B+ CTL correlated with ↓ OS.
Four et al. (2016)	32	PD-1 expression in CTLs correlated with PD-L1 expression in tumor cells. Presence of PD-1+ CTLs is associated with ↓ OS
Sasayama et al. (2016)	47	↑ number of CD68+TAMs correlates with ↓ PFS on univariate analysis but not on multivariate analysis.
Cho et al. (2017)	76	↑ expression of CD68+ TAMs is associated with ↓ OS and ↓ PFS. FoxP3 expression in Tregs has no prognostic signifcance. PD- 1 expression is associated with inferior OS
Miyasato et al. (2018)	5	PD-L1 and IDO1 were overexpressed by macrophage/microglia in PCNSL
Marcelis et. al. (2020)	36	CD8+ infiltrate and M1/M2 macropahges ratio are associated with ↑ OS. CD8+ infiltrate expresses PD-1 and TIM-3.
Alame et. al. (2021)	54	Finding of 3 immunophenotypes: rich, intermediate and poor. Rich has is a hot-like tumor expressing MHC-I, PD-L1 and STAT3. Poor and intermediate are cold-like with WNT, HIPPO and NOTCH activity.

Note:
IELSG, International Extranodal Lymphoma Study Group

Swerdlow et al., 2016; Y. Yuan et al., 2021).

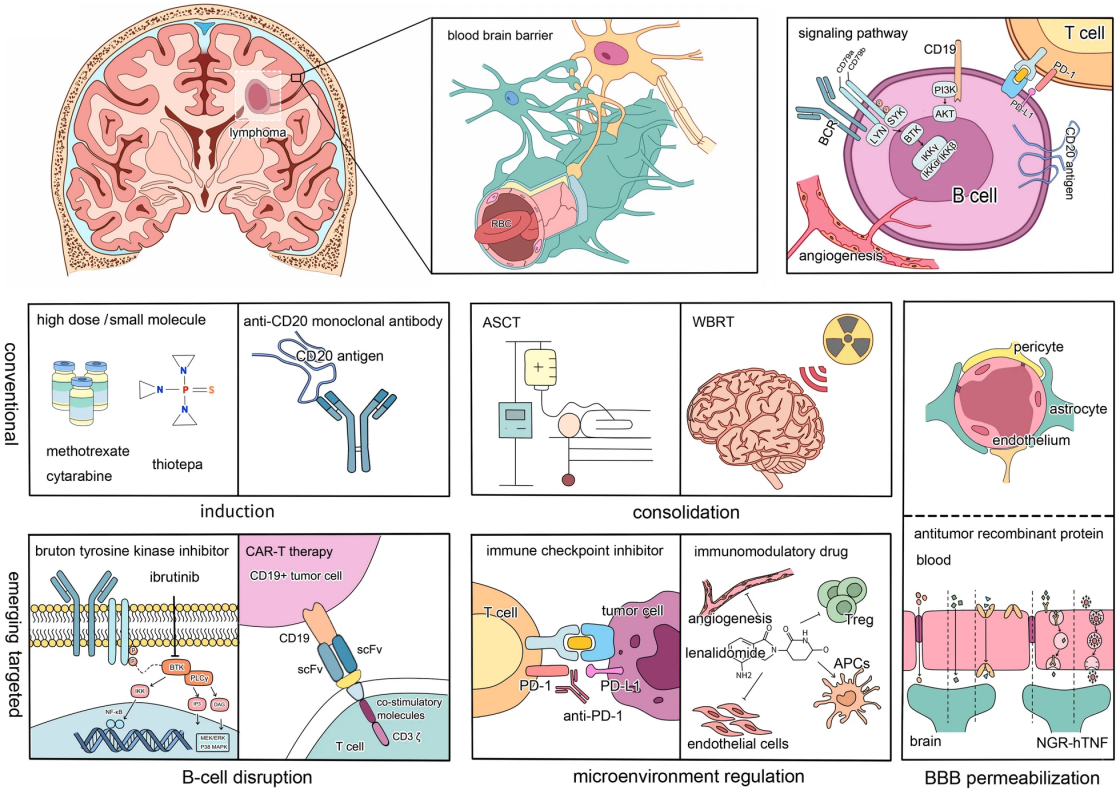


Figure 2.11: PCNSL biology, TME and therapeutic strategies. The BBB, limiting drug penetration to the tumor tissue, is formed by endothelial cells connected by tight junctions, basal lamina, pericytes, and astrocytes. Multi-signaling pathways relating to B-cell development and activation are involved. Besides HD-MTX chemotherapy during induction and ASCT or WBRT during consolidation, more specific targets in B cells and the BCR signaling pathway, immune microenvironment regulation and BBB permeabilization are currently being exploited. Adapted from Yuan et. al., 2021.

2.4.1 Prognostic factors

PCNSL has a less favorable prognosis than DLBCL as it has been reported to have 25.3 months OS with HD-MTX-based chemotherapy regimens (the current standard initial treatment) (Houillier et al., 2020). Second-line salvage therapies, including HD-MTX rechallenge, temozolomide, platine/cytarabine, topotecan, WBRT, lenalidomide, or intensive chemotherapy followed by ASCT showed heterogeneous response rates (range, 14%–85%) and survivals (range, 4–59 months) (Langner-Lemercier et al., 2016). Despite PCNSL being chemosensitive, 33% of the patients are refractory to first-line treatment, and up to 60% of the patients will eventually relapse after 2-5 years of the initial diagnosis (Houillier et al., 2020; Langner-Lemercier et al., 2016). Today there is no standard chemotherapy

salvage therapy for refractory/relapsed (R/R) PCNSL, this results in a 2.2 month (range, 0–29.6) PFS and 3.5 months (0–29.6) OS following recurrence (Houillier et al., 2020; Jahnke et al., 2006; Langner-Lemercier et al., 2016; Y. Yuan et al., 2021).

Age and Karnofsky performance score (KPS) have been long established as treatment-independent prognostic factors in PCNSL (Hoang-Xuan et al., 2015; Steven H. Swerdlow et al., 2008; Steven H. Swerdlow et al., 2016). The KPS, defined as the standard evaluation of a patient's capacities to perform an ordinary activity, ranges from 0-100 where a score of 70 is normally used as a threshold since it dictates that the patient can perform daily life activities at home but is unable to work. Today two major scoring systems, the Memorial Sloan-Kettering Cancer Center (MSKCC) and the International Extranodal Lymphoma Study Group (IELSG) prognostic models, that use age and KPS as factors are routinely used in the clinics (Table 2.5 and Table 2.7). With the evolution in the knowledge of

Table 2.5: Memorial Sloan-Kettering Cancer Center prognostic model in PCNSL. Adapted from Liu et. al. 2021.

Parameters (Age, KPS)	Median OS (years)
Age <= 50 years	8.5
Age > 50 years	3.2
Age > 50 years & KPS < 70	1.1

Table 2.7: International Extranodal Lymphoma Study Group prognostic model in PCNSL. Adapted from Liu et. al. 2021.

Parameters (each factor = 1 point)	Prognostic groups (according to the score)	2-year OS (%)
Age>60 years; ECOG	0-1	80
PS > 1; LDH > normal	2-3	48
High CSF protein level; deep brain lesions	4-5	15

Note:
ECOG, Eastern cooperative oncology group; PS, performance status; LDH, lactate dehydrogenase.

the PCNSL's molecular biology, several prognostic markers have been reported in patients with PCNSL (Table 2.9) (Cambruzzi, 2020; C. Chen, Zhuo, Wei, & Ma, 2019; I. Cho et al., 2020; Chunsong et al., 2006; Sehui Kim et al., 2019a, 2019b; Kondo et al., 2019; Le et al., 2019; Levy, DeAngelis, Filippa, Panageas, & Abrey, 2008; Makino et al., 2015; Mondello et al., 2020; Nayyar et al., 2019; Niparuck et al., 2019; Preusser et al., 2010; Tabouret et al., 2016; Takano et al., 2018; Villa et al., 2019; X. Yang et al., 2020; W. Yin et al., 2019; X.-G. Yuan et al., 2020) some of which were already mentioned in Chapter 2.3.4. Of note, potential prognostic

markers need to be adjusted by multivariate analysis (normally by the MSKCC's prognostic groupings) to avoid confounders. For example, Complete remission (CR) after HD-MTX plus ASCT has been identified as independent prognostic predictors for OS (Kondo et al., 2019).

Table 2.9: Prognostic markers associated with PCNSL. Adapted from Liu et. al. 2021.

First author, year	Prognostic markers
Favorable markers	
Makino, 2015	Completion of 3 HD-MTX cycles
Kondo, 2019	CR status at HD-MTX/ASCT
Levy, 2008	Bcl-6
Preusser, 2010	High KPS & high Bcl-6 expression
Niparuck, 2019	ECOG score <=2, multiple brain lesions, MTD <5 cm, CD10+
Chen, 2019	GLCM-homogeneity (<0.2864)
Alame, 2020	PD-1 on TILs and PD-L1 on TAMs
Cho, 2020	Serum level of soluble PD-L1 (<0.432 ng/ml)
Cambruzzi, 2020	Expression of MHC II genes, expression of bcl-6, IMO2 and CD10
Nayyar, 2019	CD79b mutations
Kim, 2019	TPD-L1- patients with a large number of CD8+ or PD-1+ TILs
Unfavorable markers	
Yuan, 2020	ECOG >3 and multifocal lesions
Tabouret, 2017	Infratentorial location and large tumor volume (>11.4 cm3)
Chunsong, 2006	CXCL13
Le, 2019	Anemia
Yang, 2020	Elevated CSF IL-10 and STAT3 phosphorylation
Yin, 2019	Bcl-2 gene aberrations and DH
Villa, 2019	Bcl-6 rearrangements
Takano, 2018	MYD88 mutations
Mondello, 2020	Tumor expression of activated STAT6, and elevated levels of IL-4 and IL-10
Kim, 2019	TPD-L1+ patients with a small number of CD8+ or PD-1+ TILs

Note:
MTD, maximum tumor diameter; GLCM, grey-level co-occurrence matrix; tPD-L1, tumoral PD-L1.

2.4.2 Induction phase

Because of the BBB impediments, the majority of the chemotherapeutics currently used to treat PCNSL are small molecules (400-600 daltons) that are capable of penetrating the BBB (Y. Yuan et al., 2021). Since the 1990s, the mainstay of treatment for patients with PCNSL is induction chemotherapy, which aims for CR, followed by consolidation therapy that aims to eradicate residual disease and improve OS (Figure 2.12) (Y. Liu, Yao, & Zhang, 2021).

HD-MTX

According to the National Comprehensive Cancer Network (NCCN) (2020) guidelines, HD-MTX chemotherapy (1–8 g/m^2) followed by WBRT consolidation ther-

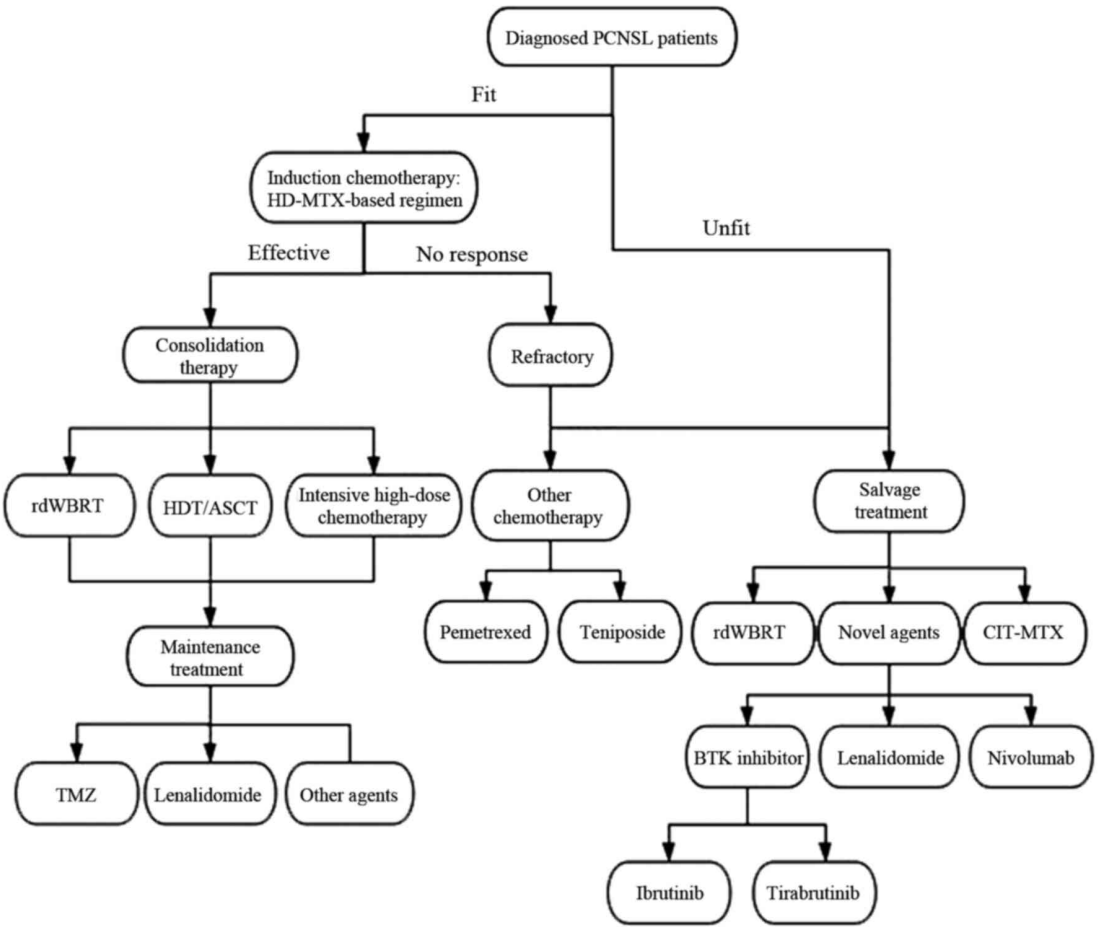


Figure 2.12: Key points in the treatment of PCNSL. rdWBRT, reduced-dose whole brain radiotherapy; TMZ, temozolomide; CIT, continuous intrathecal injection therapy; BTK, Bruton's tyrosine kinase. Taken from Liu et. al., 2021.

apy, is the systemic therapy for newly diagnosed patients with PCNSL. Concerning the optimum dose of methotrexate, although there is no clear evidence of a dose-response, a dose of 3 g/m^2 or above in a rapid infusion is recommended. MTX is an antifolate that suppresses DNA synthesis by inhibiting dihydrofolate reductase activity in purine and thymidine synthesis, which controls the expression of glucocorticoid receptors in blood cells (Houillier et al., 2020; Ricard et al., 2012; Y. Yuan et al., 2021). Moreover, based on disappointing experiences with HD-MTX as a single agent, polychemotherapy is recommended since it offers a better prognosis. Polychemotherapy includes HD-MTX combined with prednisone, vincristine and temozolomide (TMZ). Additionally, while HD-MTX-based treatment is widely accepted in clinical settings, 50% of patients may have a risk of progression or recurrence (Y. Liu, Yao, & Zhang, 2021).

2.4.3 Consolidation phase

Because even after intensive MTX treatment a total of 20-30% of patients with PCNSL relapse within 6 months, consolidation therapy is required to eliminate minimal residual disease. Consolidation options include WBRT, HDT/ASCT, and intensive HD chemotherapy (Y. Liu, Yao, & Zhang, 2021).

2.4.4 Maintenance treatment

Maintenance treatment serves as an alternative approach to prolonging remission, delaying relapses, and maintaining tumor dormancy when patients wish to avoid WBRT neurotoxicity or cannot tolerate consolidation therapies. Options include agents, such as TMZ, procarbazine, lenalidomide and ibrutinib (Y. Liu, Yao, & Zhang, 2021).

2.4.5 Salvage treatment

Even with the advances in the induction and consolidation options, 10-15% of patients with PCNSL become refractory to initial treatment, and 35-60% relapse within 1-2 years (Hoang-Xuan et al., 2015; Y. Liu, Yao, & Zhang, 2021). Salvage treatments for R/R PCNSL patients depend on age, KPS, site of relapse within the CNS, prior treatments, and time duration from last response (Hoang-Xuan et al., 2015). Options include HDT/ASCT, chimeric antigen receptor T-cell (CART-cell) therapy, BTKis, and lenalidomide.

2.4.6 Other treatments

Thanks to recent molecular information other potential treatments are under ongoing clinical trials, such as ICIs, PI3K/AKT/mTOR signaling inhibitors, and cyclin-dependent kinase (CDK) 4/6 inhibitors.

2.4.7 Future perspectives

Understanding PCNSL biology has allowed incorporating target therapies and immunotherapies, but unfortunately, heterogenous responses are still observed for most of the current treatment strategies. Therefore, identifying subgroups of PCNSL patients with shared causative biologic factors of disease and outcome is of extreme importance to tailor treatment strategies. However, such molecular subgroups' identification is extremely challenging mainly due to high genetic, phenotypic, and TME heterogeneity. Consequently, there is an unmet need to perform large multi-omics studies with a view to personalizing clinical care and to improving patients' outcomes.

Chapter 3

Multi-omic era for molecular subtyping

Thanks to the advent of NGS technology, collection of data from different molecular compartments (e.g. genetic alterations, gene expression, DNA methylation status, protein abundance) have been possible, ultimately resulting in multiple omics (multi-omics) data obtainment with a view to a comprehensive understanding of cancer (Huang, Chaudhary, & Garmire, 2017). The addition of "omics" to a molecular term implies a comprehensive, or global, assessment of a set of molecules; for example, transcriptomics refers to the study of RNA levels genome-wide, both qualitatively (which transcripts are present, identification of novel splice sites, RNA editing sites) and quantitatively (how much of each transcript is expressed) (Hasin, Seldin, & Lusis, 2017). Bioinformatic analyses of single-omic data have increased disease comprehension, which is the case of the PCNSL molecular definitions described in Chapter 2.2; however, this approach often misses the complexity of the landscape of molecular phenomena underlying the disease. On the other hand, multi-omic approaches consider interactions between omics layers, hence providing a more accurate reconstruction of molecular networks. Both single- and multi-omics approaches have the ultimate goal of finding biomarker signatures of a specific disease, discovering disease subtypes, and predicting response to therapy or survival time (Momeni, Hassanzadeh, Saniee Abadeh, & Bellazzi, 2020). Therefore, recent cancer research has been eager to build more complex models by means of multi-omic data analysis; nevertheless, this comes with three major challenges which are acquisition, reduction, and integration of the multi-omic data (Figure 3.1).

This chapter aims to present an overview of the main bioinformatic concepts and tools needed for acquiring, analyzing, and applying multi-omic data in the context of cancer research. Subchapter 3.1 covers which are the most relevant features to extract from each single-omic data (e.g. genomics, epigenomics, transcriptomics, clinicOmics) but only focusing on the omic data used for this thesis. In subchapter 3.2, methods for reducing single-omic's features are presented. Finally, subchapter 3.3 focuses on how multi-omic data integration can be used to answer biological questions that have clinical relevance.

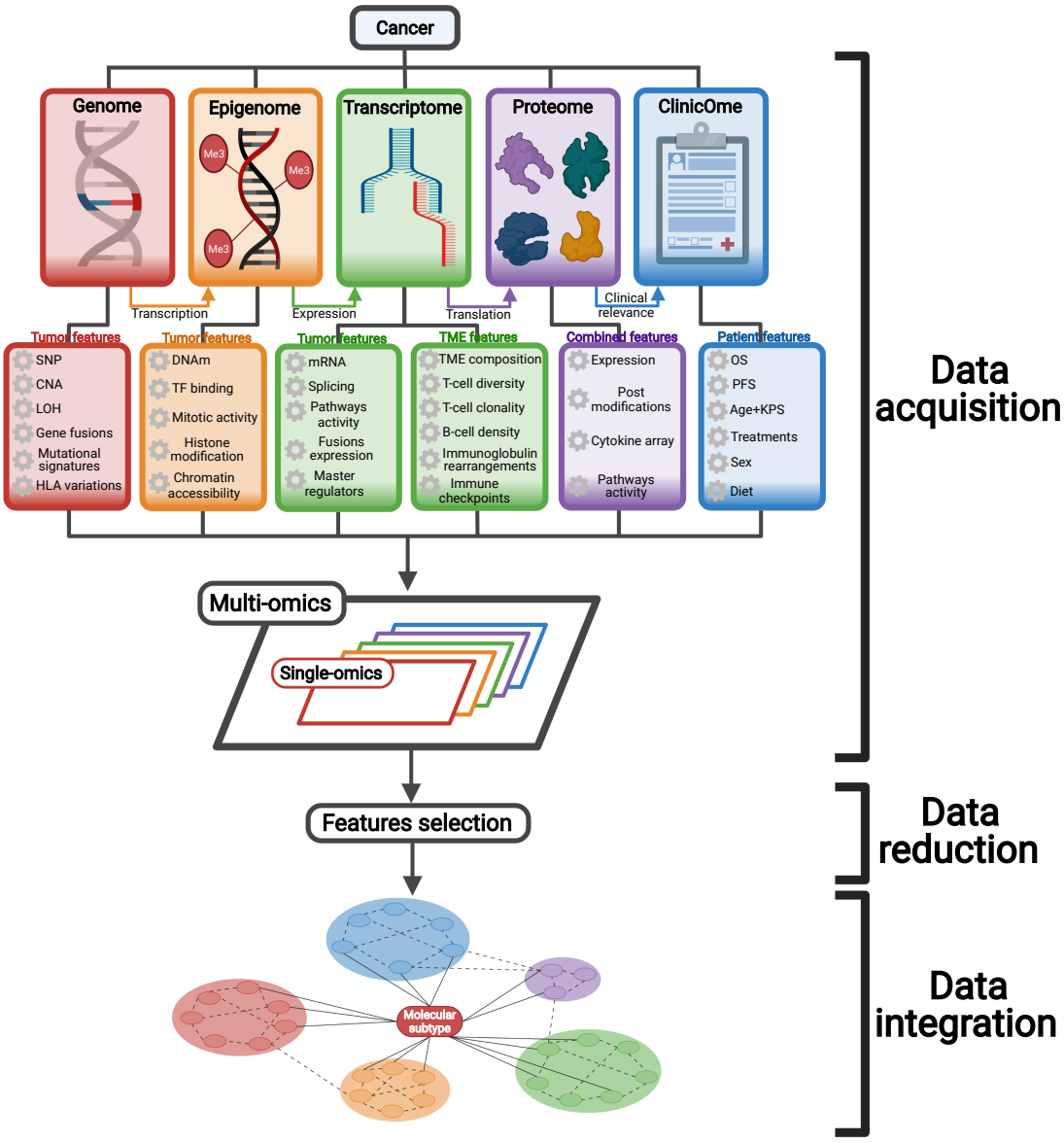


Figure 3.1: Illustrative diagram of multi-omic data analysis for molecular subtyping. Different layers of single-omic data have to be processed to acquire various features which can be related to either the tumor, the TME, or the patient. The result of combining all the features from these layers is termed "multi-omics". As thousands of features are recovered, the next challenge is to subset those that can effectively describe the system and its complex interactions. Finally, multi-omic data integration, although challenging, can be used to find molecular subtypes of disease with shared causative biologic factors.

3.1 Adquisition of multi-omic data

Prior to acquiring any single-omic layer of information, a lot of collaborative and organized effort has to take place to recover enough genetic material from each patient comprising a cohort. Such effort includes the study design, ethical approval, patients' diagnosis, biopsies extraction, the signature of informed consents, sample preservation, among others. Afterwards, depending on the nature of the omic information different parameters have to be considered to firstly, obtain high-quality raw data (eg. raw sequencing files), and secondly, extract distinct features from that data. Here, a feature is defined as a distinctive attribute that can be extracted from an omic data, for example, information concerning single nucleotide variants (SNVs), CNAs, gene fusions, mutational signatures, and HLA-variations can be indirectly obtained from genomic data by means of whole-exome sequencing (WES) or whole-genome sequencing (WGS) bioinformatic analysis. Moreover, the nature of the omic information also dictates the origin of the features that can be extracted; for example, transcriptomic information can render both tumor and TME related features (Figure 3.1). Features from different single-omic layers can also be combined to extract additional information; for example, the SNPs and CNA from the genomics can be combined with the MHC expression from the transcriptomics to obtain a list of neoantigens that can possibly be present at the surface of the cancer cell (Schmidt et al., 2021). This section covers the definition of the different types of features that can be extracted from the genomic, epigenomic, transcriptomic, and clinicomic layers.

3.1.1 Genomic data

A vast majority of cancer genomes contain many nucleotide sequence changes compared with the germline of the cancer patient. Some of such variations can cause or promote cancer, referred colloquially as "drivers," while others seem not to present obvious advantages to the cancerous cells, those are referred to as "passengers." These alterations give rise to genomic tumoral features such as SNVs, insertions/deletions (indels), CNAs, gene fusions, that can be produced by different mutational processes (e.g. ultraviolet radiation) (Australian Pancreatic Cancer Genome Initiative et al., 2013; Chin, Hahn, Getz, & Meyerson, 2011; PCAWG Mutational Signatures Working Group et al., 2020).

SNV, indels & CNA

WES or WGS generate raw files named FASTQ which are a text-based format for storing nucleotide next-generation sequence reads and their corresponding per-base quality scores as well as information relating to whether reads are single-end or paired-end. Those raw reads need to be mapped to the human reference genome, from which various versions exist,using, for example, the BWA-MEM software (H. Li & Durbin, 2009). Sequence alignment/map (SAM) or BAM (binary version of SAM) files result from this alignment and contain details of aligned and unaligned

reads are stored along with associated mapping qualities. Then, there is a rigorous process of data cleanup that includes marking PCR duplicates, realigning indels, and recalibrating bases (reduces mismatches, degradation, and sequencing artifacts); the tools needed for this process are integrated into the Genome Analysis Toolkit (GATK) software (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). Resulting "cleaned BAMs" are subsequently used for the variant discovery phase after which a variant call format (VCF) file, containing somatic mutations along with read depths for the reference and alternate alleles, is produced. VCFs can be produced by only one calling algorithm (e.g. Mutect); however, a combination of at least 3 (e.g. Mutect2, Strelka, VarScan, Lofreq, SomaticSniper) is preferred to reduce false-positive rates (Fang et al., 2015; Sangtae Kim et al., 2018; Koboldt et al., 2012; Larson et al., 2012; McKenna et al., 2010; Wilm et al., 2012). Of note, during the variant calling process, germline variants need to be filtered out by using either sequencing files from a matched normal counterpart (normally DNA from blood) or a panel of normals (PON) when only the tumoral sequences are available (Van der Auwera et al., 2013). Next, the SNVs and indels contained with a VCF file are annotated, this means adding meta-information about the likely effect on genes, transcripts, and protein sequence. Annotation, which generates a mutation annotated format (MAF) file, cab be achieved using Variant Effect Predictor (VEP) program from Ensembl (McLaren et al., 2016) (Figure 3.2).

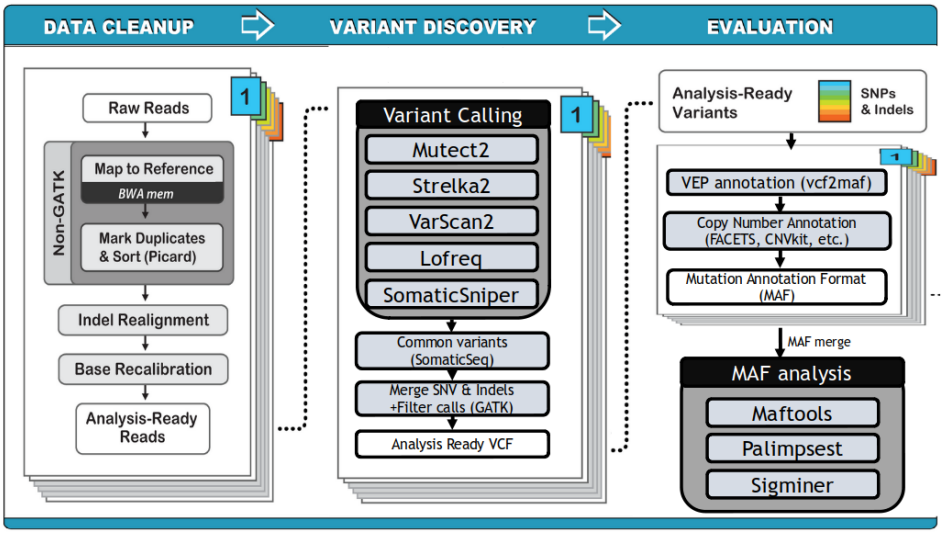
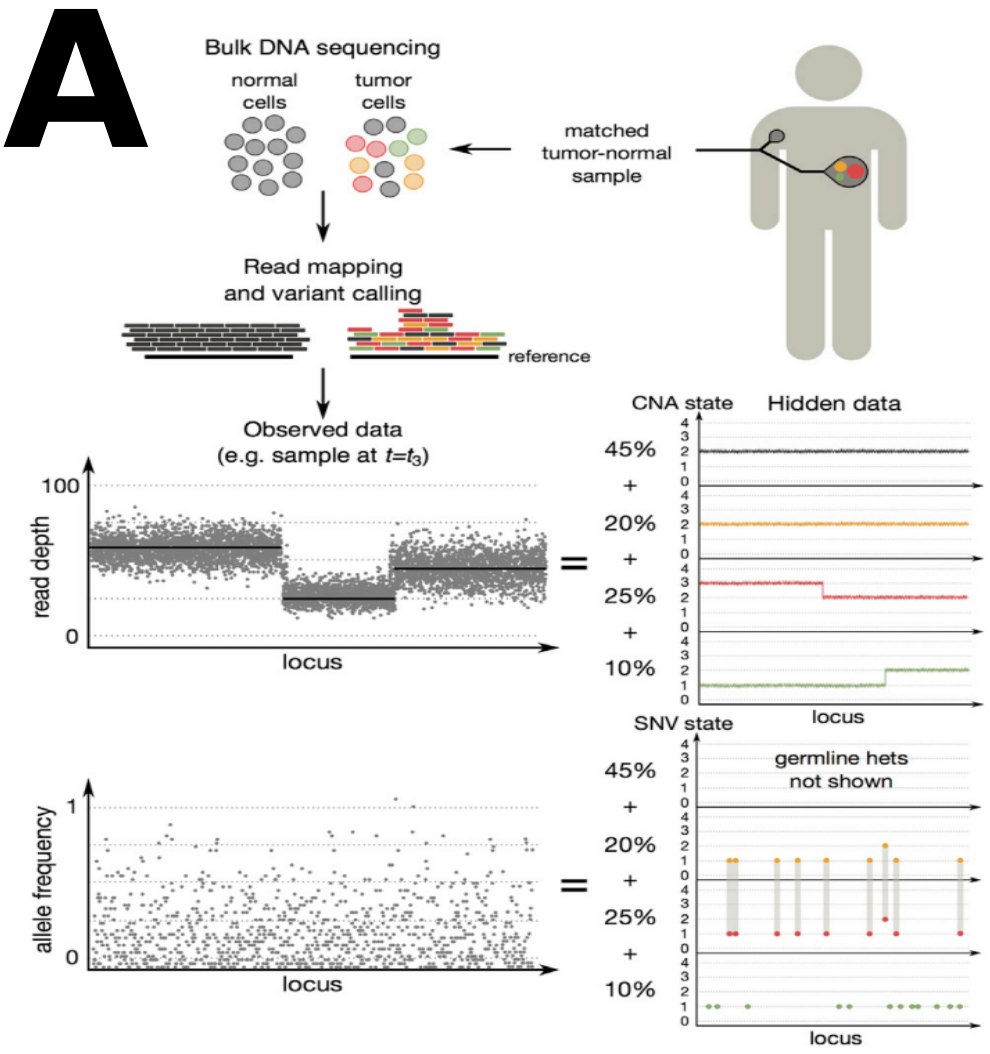


Figure 3.2: MAF generation workflow.

Besides SNPs and indels, CNA can also be retrieved using "cleaned BAMs" as input and the help of software like FACETS, CNVkit, copynumber, among others (Nilsen et al., 2012; Shen & Seshan, 2016; Talevich, Shain, Botton, & Bastian, 2016). The CNAs found in cancer include whole-chromosome or regional alterations spanning part to whole arms of chromosomes, as well as focal events involving one or a few genes. Moreover, one can infer cancer clonal heterogeneity by

using CNAs, read depths, B-allele fractions (BAFs), and SNVs coupled with hidden Markov models. Clonal heterogeneity is given by the presence of a collection of subclones within the fraction of cancerous cells where each subclone population presents a set of private and shared mutations related by their joint evolutionary history going back to the most recent common ancestor (Fischer, Vázquez-García, Illingworth, & Mustonen, 2014). Determining mutations' clonality has important biological and medical implications; for example, one can infer driver mutations arising at very early stages or late mutations being acquired after disease recurrence (PCAWG Evolution & Heterogeneity Working Group et al., 2020; Touat et al., 2020). Softwares for inferring clonality include MutationTimer, Palimpsest, and cloneHD (Fischer, Vázquez-García, Illingworth, & Mustonen, 2014; PCAWG Evolution & Heterogeneity Working Group et al., 2020; Shinde et al., 2018) (Figure 3.3).



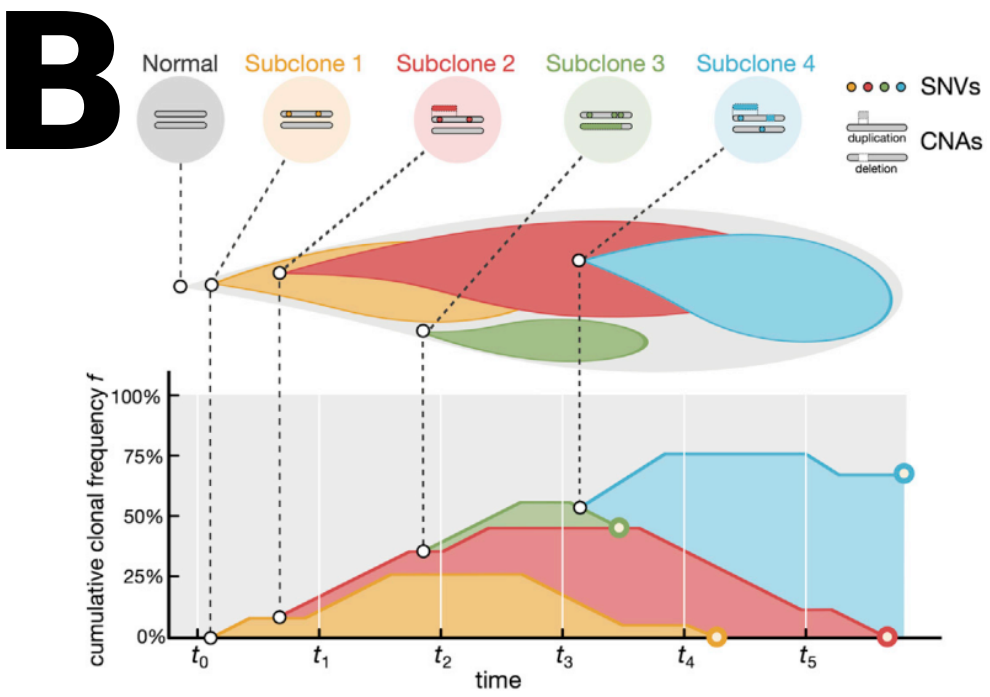


Figure 3.3: Clonal heterogeneity reconstruction. Panel A shows the schematic illustration of the CNA and SNV information needed for inferring clonality. Panel B shows a schematic view of subclonal diversification where a point mutation occurs early on with a subsequent gain of a chromosome arm and a short deletion at a later stage, each followed by clonal expansion (subclones 1, 2, and 4). A short-lived lineage arises independently (subclone 3). Adapted from Fischer et. al., 2014.

Gene fusions

Gene fusions are genomic rearrangements leading to structural translocations, chromosomal inversions, or interstitial deletions that can have an important clinical impact (e.g. *BCL6*, *BLCL2*, and *IG* translocations) (Chapuy et al., 2016). Detecting gene fusions is challenging mainly because they resemble common sequencing and alignment artefacts. Typically, they are detected by looking for changes in read depth, identifying clusters of discordantly aligned paired-end reads or split read, constructing some form of assembly or a combination of these approaches. Some callers include Pindel, BreakDancer, manta, BreakPointer, among others (Cameron, Di Stefano, & Papenfuss, 2019). Moreover, even though fusions can be formed during cancer progression their expression is not always activated, hence complementary programs for detecting their expression are highly recommended (Uhrig et al., 2021).

Mutational signatures

By looking at the trinucleotide context flanking somatic mutations (96 possible combinations called mutation spectra), recent research has suggested the existence of different mutational processes (also called signatures) of both exogenous (i.e. UV light exposure, smoking, chemotherapy) and endogenous (i.e. DNA damage, DNA repair/replication deficiencies) origin. Initially, 21 distinct validated mutational signatures were described in 2013; for example, C>T substitutions at NpCpG trinucleotides was related to spontaneous deamination of 5-methyl-cytosine (age-related, Figure 3.4) (Australian Pancreatic Cancer Genome Initiative et al., 2013). Seven years later, Alexandrov et. al. extended the single base substitution (SBS) signature catalog to 96 by using two software based on nonnegative matrix factorization (NMF) SigProfiler and SignatureAnalyzer (a Bayesian variant of NMF) (Bergstrom, Barnes, Martincorena, & Alexandrov, 2020; PCAWG Mutational Signatures Working Group et al., 2020). NMF receives the matrix of mutation spectra as input to determine the signature profiles and contributions of each signature to each cancer genome; however, the mutations observed in a particular sample can be reconstructed in multiple ways due to signatures' overfitting/underfitting or high heterogeneity in the cohort (Maura et al., 2019). Interestingly, those mutational signatures leave imprints associated with the biological/epidemiological origin of a particular cancer type; for example, while malignant melanoma is characterized by the presence of the signature SBS7 (UV related), DLBCL shows a high SBS9 signature which is related to non-canonical AID activity (Chapuy et al., 2018; PCAWG Mutational Signatures Working Group et al., 2020). Additionally, combining clonality information with mutational signatures have been recently used to determine the relative time when a mutational process is more or less active; for example, in chronic lymphoid leukaemia (CLL) the signature SBS9 is 20 times more present at early times (clonal) than at late times (subclonal) (PCAWG Evolution & Heterogeneity Working Group et al., 2020).

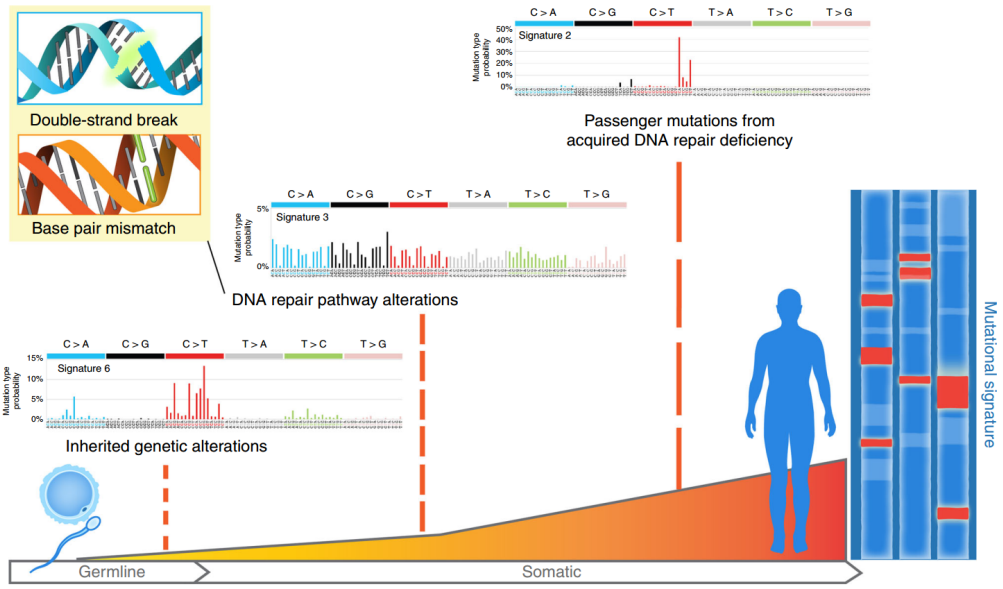


Figure 3.4: Mutational signatures. As time passes, DNA alterations, such as carcinogens or DNA repair pathway defects, leave fingerprints reflected in an individual's mutational spectra. Adapted from Ma et. al., 2018.

Neoantigens production

As previously discussed in Chapter 2.3.2, neoantigens production and presentation to CD8+ T cells are affected by many factors; however, from the bioinformatic's perspective, it is reduced to programs calculating the pHLA complex binding affinity and those calculating the TCR affinity to the pHLA complex. Generally speaking, programs like pVACtools, MuPeXI, and TIminer, take a VCF, gene expression estimates (from RNA-seq data), and HLA haplotypes (from RNA-seq or WES data, depending on the method) to compute, firstly, a list of possible 8- to 12-mer peptide list for each mutation contained within the VCF; and secondly, the MHC I affinity for each of the previously calculated peptide sequences (Figure 3.5) (Boegel, Castle, Kodysh, O'Donnell, & Rubinsteyn, 2019; Hundal et al., 2020). On the other hand, TCR affinity to the pHLA complex has been so far only implemented by PRIME which takes the neoantigen's peptide sequence and the HLA haplotypes as the input. PRIME calculates immunogenicity (i.e. the ability of a molecule or substance to provoke an immune response) by computing a %rank score which is the fraction of random 700,000 8- to 14-mers that would have a score higher than the peptide provided in input. Neoantigens are classified as "Immunogenic" if having a PRIME %rank score lower or equal to 0.5% for the corresponding HLA haplotype of the patient where the neoantigen occurred, or as "Non-Immunogenic" otherwise (Schmidt et al., 2021). Moreover, neoantigens data can also be combined with clonality and mutational signatures to give insights into biological processes or genes producing highly immunogenic neoantigens.

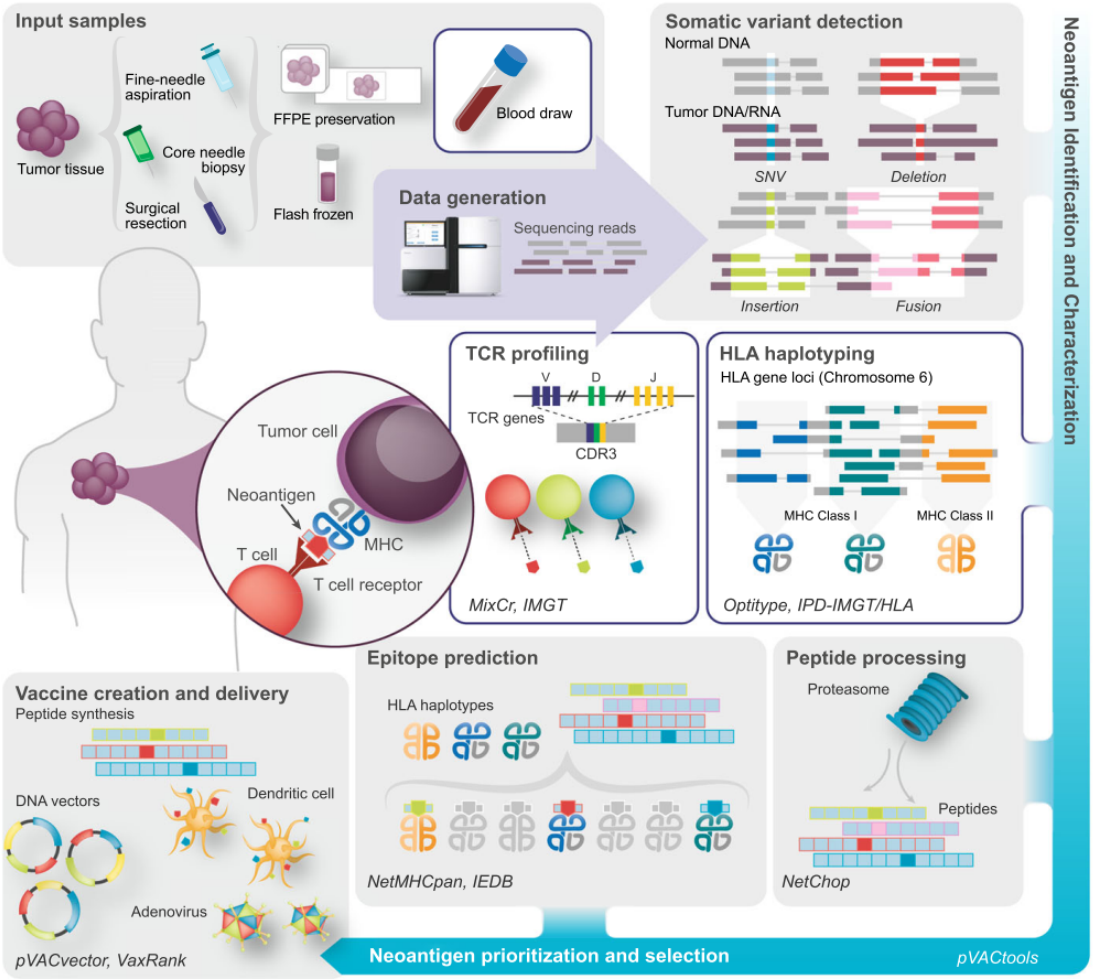


Figure 3.5: Bioinformatic overview of neoantigens' characterization. Patient sequences are analyzed to determine human HLA haplotypes, somatic variants, and their corresponding RNA expression. Next, the corresponding peptide sequences are analyzed concerning their predicted expression, processing, and ability to bind the patient's MHC complexes. Candidates are then selected for vaccine design and additional analyses are performed to assess the T-cell response. Specific exemplar bioinformatics tools for each step are indicated in italics. Taken from Richters et. al., 2019. IEDB, Immune Epitope Database.

3.1.2 Epigenomic data

Generally speaking, epigenetics acts through two mechanisms: (1) modifications to chromosomal proteins that alter the 3D conformation of the genome and/or protein-DNA interactions and (2) chemical modification of the DNA strand itself. The first mechanism can lead to either tightly packed and inactive conformations or open and accessible DNA (termed heterochromatin and euchromatin respectively); whereas the second, is the methylation of cytosine to 5-methylcytosine (5mc) at CpG sites through the action of the DNA methyltransferase enzymes

(DNMTs) (Locke et al., 2019). Cytosine methylation of DNA is an important epigenetic mechanism to control gene expression, silencing, genomic imprinting, cancer development, and regulation of the immune system. Today, WGBS constitutes the current gold standard for DNA methylation profiling due to its genome-wide coverage and single-basepair resolution. WGBS, like most DNA methylation profiling assays, rely on bisulfite treatment to selectively convert unmethylated cytosines (including 5-formyl-cytosine and 5-carboxy-cytosine) into uracil (which is subsequently replaced by thymine) while leaving methylated cytosines unconverted (Krueger & Andrews, 2011; Müller et al., 2019). Within the normal context, most CpG sequences in the genome are methylated, but CpG islands and the nearby CpG island shores (the region within 2 kb of the islands) are exceptionally hypomethylated. Furthermore, it has been reported that CpG methylation can directly repress transcription by preventing binding of some TFs to their recognition motifs which are frequently observed within tumor suppressor genes (TSGs). Conversely, gene bodies of highly expressed genes are heavily methylated (Nishiyama & Nakanishi, 2021; Y. Yin et al., 2017) (Figure 3.6).

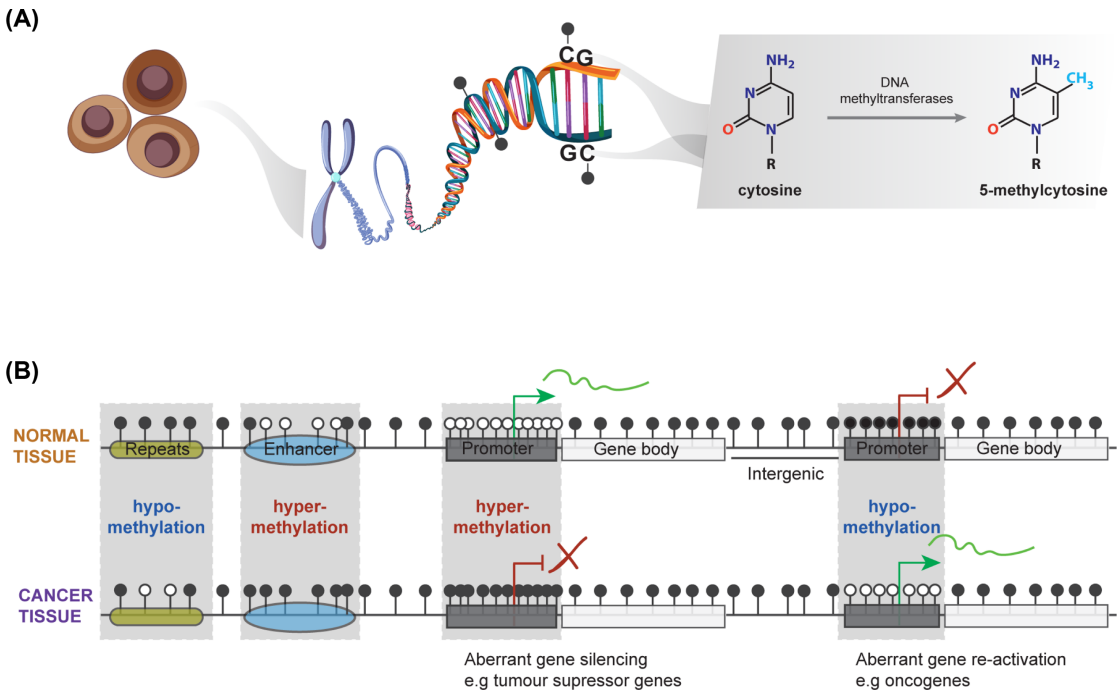


Figure 3.6: DNA methylation profiles within the normal and the tumoral contexts. DNMTs catalyze the addition of a methyl group to the fifth carbon position of cytosines primarily within CpG contexts (Panel A). Tumorigenesis drives changes in DNAm distribution causing hypermethylation of tumor suppressor genes and hypomethylation of oncogenes (Panel B). Taken from Skvortsova et. al., 2019.

Differential DNA methylation

Differential methylation (DM) analysis starts by aligning FASTQ files to the reference human genome using the software Bismark to generate BAM files which are subsequently used to exclude any duplicate calls from overlapping read ends of short inserts (Krueger & Andrews, 2011). After obtaining coverage files, DM analysis can be performed on genes, promoters (defined as the 1,500 bases upstream and 500 bases downstream of the transcription start sites of corresponding genes), CpG islands or regions. The most commonly used software for DM include RnBeads, Methrix, bsseq, and methylKit (Mayakonda et al., 2021; Müller et al., 2019; Y.-H. Zhou, Xia, & Wright, 2011).

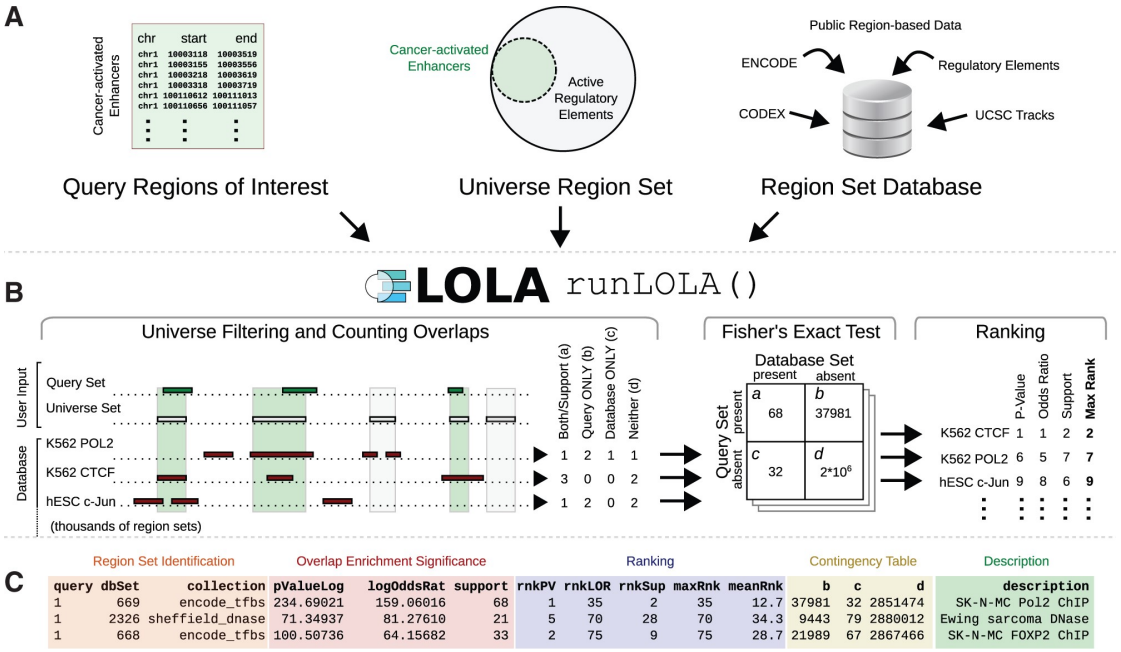
TF binding

DM results are normally divided as hypermethylated or hypomethylated for subsequent downstream enrichment analysis. For example, DM promoters can be used as input for the R-based program Locus Overlap enrichment Analysis (LOLA) which computes the enrichment of the input data against reference Transcription factor binding sites (TFBS) databases. LOLA analysis comprises three components, the query set—one or more lists of genomic regions to be tested, a region universe—the background set of regions that could potentially have been included in the query set, and a reference database of genomic region sets that are to be tested for overlap with the query set (Figure 3.7 A). LOLA evaluates enrichment using Fisher's exact test with false discovery rate correction to assess the significance of overlap in each pairwise comparison; next, it ranks each region set using the p-value, log odds ratio, and the number of overlapping regions (Figure 3.7 B and C). Moreover, this same principle can be further applied to other databases such as polycomb-associated zones (Sheffield & Bock, 2016).

Mitotic activity

As stem cell divisions and populations increase within a tissue, so does the chronological age of an individual and the error in the maintenance of DNAm. In addition, increased mitotic rate due to cancer risk factors such as inflammation or viral infection has been suggested to fuel epigenetic cellular heterogeneity and to lead to increased epigenetic activity. In the light of these findings, efforts have been done to construct mitotic clocks based on DNAm, for example, Horvath constructed an epigenetic clock that uses 353 CpG sites associated with chromatin states and tissue variance while Yang et. al. developed epiTOC to demonstrate that methylation is universally accelerated in cancer (Horvath, 2013; Z. Yang et al., 2016). More specifically in the context of B-cells, Duran-Ferrer et. al. developed a mitotic clock, called epiCMIT, that represents a relative measure of the total proliferative history of normal and neoplastic B-cells (Duran-Ferrer et al., 2020).

epiCMIT is an R-based program that takes a DNAm matrix as input to return two underlying hyper- and hypomethylation-based mitotic clocks (called epiCMIT-hyper and the epiCMIT-hypo, respectively), and a last one which is the highest



query	dbSet	collection	pValueLog	logOddsRat	support	rnkPV	rnkLOR	rnkSup	maxRnk	meanRnk	b	c	d	description
1	669	encode_tfbs	234.69021	159.06016	68	1	35	2	35	12.7	37981	32	2851474	SK-N-MC Pol2 ChIP
1	2326	sheffield_dnase	71.34937	81.27610	21	5	70	28	70	34.3	9443	79	2880012	Ewing sarcoma DNase
1	668	encode_tfbs	100.50736	64.15682	33	2	75	9	75	28.7	21989	67	2867466	SK-N-MC FOXP2 ChIP

Figure 3.7: LOLA workflow. Example of LOLA input requirements (A), calculations (B), and enrichment results (C). Taken from Sheffield and Bock, 2016.

value between the other two (called epicMIT). All of them range from 0 to 1, depending on low or high relative proliferative history (Duran-Ferrer et al., 2020).

3.1.3 Transcriptomic data

The ~ 22,000 protein-coding genes and many regulator elements in the human body, comprised within the transcriptome, add an extra layer of complexity for understanding cancer. Over the past few decades, transcriptome profiling has evolved from microarrays to high-throughput sequencing at different resolution levels such as bulk RNA sequencing (RNA-seq), scRNA-seq, and spatial RNA-seq. As a versatile tool, bulk RNA-seq can retrieve features from both the tumor and the TME such as RNA expression (gene, transcripts, exons, or fusions), master regulators' and specific pathways' activity, TME composition, neoantigens, and TCR/BCR clonotypes (Figure 3.1).

mRNA expression

Similar to WES/WGS data processing, RNA-seq produces FASTQ files which are submitted to alignment after they passed quality control using programs such as FASTQC (preferred) and NGSQC (Andrews, 2010; Patel & Jain, 2012). The next step is to trim adapter sequences, which can greatly impact alignment efficiency, using programs like Trimmomatic, CutAdapt, or PRINSEQ (Bolger, Lohse, & Usadel, 2014; Martin, 2011; Schmieder & Edwards, 2011). Adapter trimming de-

pends on the sequencing technology used, for example, poly-A tails need to be trimmed when using 3' RNA-seq (Martin, 2011). Cleaned FASTQ files are then aligned to the reference transcriptome to produce BAM files, where, depending on the desired information, there are distinct programs and parameters to be used. For example, the STAR software allows FASTQ mapping with subsequent read-counts quantification for genes, transcripts, and even gene fusions (Dobin et al., 2013). On the other hand, DEXSeq is a software focused on exon level quantification (Anders, Reyes, & Huber, 2012), while tools such as Kallisto, Sailfish, and Salmon are some alignment-free quantification tools (they do not produce BAM files) (Bray, Pimentel, Melsted, & Pachter, 2016; Patro, Duggal, Love, Irizarry, & Kingsford, 2017; Patro, Mount, & Kingsford, 2014). Moreover, only the information stored in BAMs from pair-ended sequencing data can be properly used by downstream analysis programs (e.g. MixCR, immunearch, arriba) to obtained additional features such as TCR/BCR clonotypes, gene fusions expression, splicing isoforms, among others (Anders, Reyes, & Huber, 2012; Bolotin et al., 2015; ImmunoMind Team, 2019; Uhrig et al., 2021). Another aspect to take into account is tissue origin since it is not the same to align reads coming from "intact" FF tissue as those coming from "degraded" FFPE origin. DegNorm is a recently developed program that can correct the degradation contribution due to the paraffin fixation process (Xiong, Yang, Fineis, & Wang, 2019). Next, raw counts are used as input for generating a normalized expression matrix and finding differentially expressed genes (DEGs) (Love, Huber, & Anders, 2014; M. D. Robinson, McCarthy, & Smyth, 2010). Finally, the DEGs can be used to discover potential cancer theranostic biomarkers. An overview of these bioinformatic tools and steps is illustrated in Figure 3.8 (Hong et al., 2020).

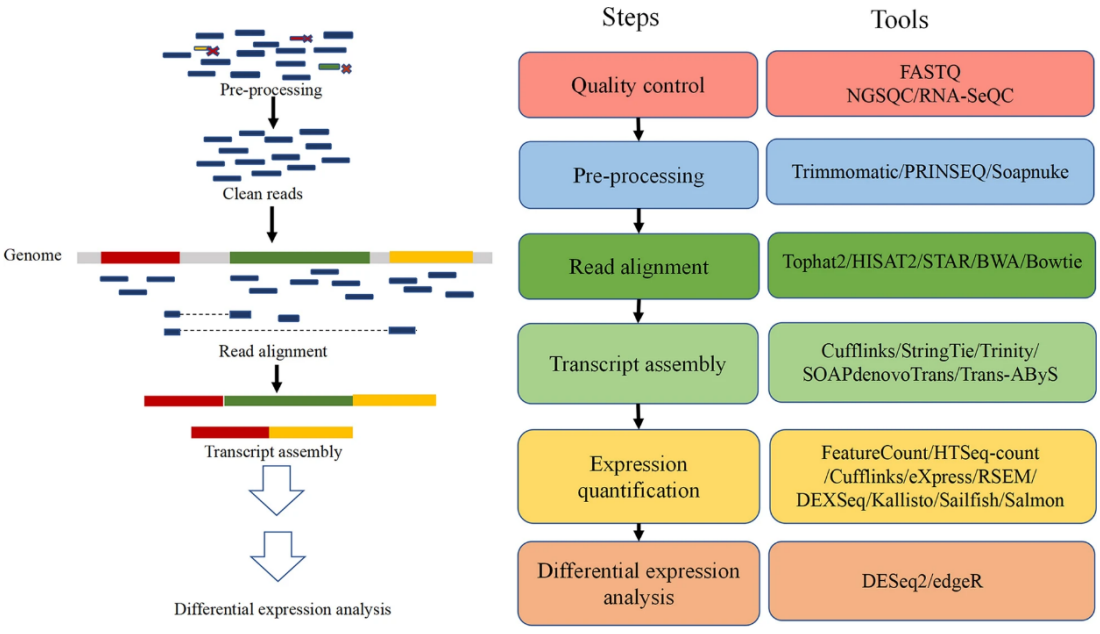


Figure 3.8: Bioinformatics tools commonly used in RNA-seq data analysis. Taken from Hong et. al., 2020.

Gene fusions expression

Gene fusions play a major role as oncogenic drivers in many cancer types and as therapeutic targets (Uhrig et al., 2021). Although gene fusions can be detected as SVs using WES/WGS data, RNA/protein expression corroboration is normally preferred. However, reliable prediction of gene fusions from short-read RNA-seq can be difficult mainly owing to a myriad of artifacts being introduced during library preparation and sequence alignment. Algorithms to detect gene fusions include Arriba (preferred), STAR-Fusion, pizzly (compatible with kallisto pseudo alignments), and defuse (Haas et al., 2019; McPherson et al., 2011; Melsted et al., 2017; Uhrig et al., 2021).

Pathways activity

Despite differential gene expression analysis being one of the most common applications of RNA-seq, gaining insights into the biological processes underlying phenotypic differences can become difficult when looking at individual genes. gene set analysis (GSA) methods use DEGs and incorporate pre-existing biological knowledge (in a form of functionally related gene sets or known biological pathways) to gain insights into molecular processes. Among GSA methods, gene set overrepresentation analysis is widely used with the help of annotated gene sets such as Gene Ontology (GO) or Kyoto Encyclopedia of genes and genomes (KEGG), using standard statistical tests for enrichment (e.g. Fisher test); however, it does not account for genes with small changes in expression that might be biologically relevant. Alternative techniques that consider the differential expression of gene sets and do not require a priori selected genes include Gene Set Enrichment Analysis (GSEA), single sample Gene Set Enrichment Analysis (ssGSEA), and Gene Set Variation Analysis (GSVA). GSEA tests the null hypothesis that the genes (normally the DEGs) in a gene set (manually curated or from databases) are randomly associated with the phenotype by performing weighted Kolmogorov–Smirnov statistics. On the other hand, the GSVA and ssGSEA methods can be applied to normalized gene expression matrices to calculate sample-wise enrichment scores where the statistics behind these calculations differ. GSVA compares the cumulative distribution function, resulting from Kolmogorov rank statistics, of all the genes in a gene set from Sample_i versus the empirical distribution resulting from all the samples within the dataset. ssGSEA sample-wise enrichment score is calculated as a sum of the differences between two weighted empirical cumulative distribution functions of gene expressions inside and outside the gene set (Barbie et al., 2009; Hänzelmann, Castelo, & Guinney, 2013; Rahmatallah, Emmert-Streib, & Glazko, 2016; Subramanian et al., 2005).

Additionally, since the major difference between various GSA approaches remains in the null hypothesis they test and is unaffected by the data type being used, they can be used with other data sources like DNAm, protein expression, or lists of genetic alterations. Software to perform these analyses include R-based GSVA, clusterProfileR, webgestalt, and enrichR (Hänzelmann, Castelo, & Guin-

ney, 2013; Kuleshov et al., 2016; G. Yu, Wang, Han, & He, 2012; B. Zhang, Kirov, & Snoddy, 2005).

TFs and master regulators (MRs)

In addition to epigenetic control, gene expression programs, which are fundamental for cell development, differentiation, tissue homeostasis, and disease, can be further regulated by TFs through their interaction with specific DNA regulatory regions (Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019). A TF regulates the activity of a collection of target genes (the so-called regulon) and, within each TF network, such regulons may overlap with those of other TF since they can be regulated by different "branches." Furthermore, within a specific pathway, TFs can be regulated by other TFs which receive the name of master regulators if they are at the top of a gene regulation hierarchy. Consequently, given the biological importance of MRs and TFs and because high-throughput measurements of their activities are not available, various computational and biological approaches have been constructed to estimate their activities from the gene expression levels of their regulons (Fletcher et al., 2013; Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019). From a practical bioinformatics point of view, two R-based software (DoRothEA and RTN), which are complementary, are commonly used to estimate TFs/MRs activity using normalized gene expression matrices as input (M. A. A. Castro et al., 2016; Garcia-Alonso, Holland, Ibrahim, Turei, & Saez-Rodriguez, 2019).

DoRothEA uses the Wald statistic results from the DEGs (retrieved from running DESeq2), retrieved human TF–target interactions for 1,541 TFs (from experimental and literature-curated resources), and statistical methods from virtual inference of protein activity by enriched regulon analysis (VIPER) R-package to compute TF regulon normalized enrichment scores (NESs). Briefly, unlike other GSA methods like GSEA, Fisher's exact test, GSVA, VIPER incorporates directionality by integrating different likelihoods of representing activated, repressed, or undetermined targets and probabilistic weighting of low vs. high-likelihood protein targets (Alvarez et al., 2016) (See Figure 3.9 A&B).

Unlike DoRothEA, the RTN package comes with a less extensive list of TF–target interactions but provides better graphical outputs by internally using the RedeR package and provides MRs' activity calculation. Regulator-target associations are identified using: I) mutual information (MI) which indicates whether or not a regulator is well informative of the status of a target gene and II) the direction of the association (positive or negative) evaluated by Spearman's correlation. Furthermore, associations can be filtered first by retaining only edges with a BH-adjusted P-value < 0.01 (recommended but can be adjusted) after permuting the MI matrix 1000 times; secondly by eliminating unstable interactions by 1000 times resamples' bootstrapping (consensus bootstrap $> 95\%$); and finally by removing indirect TF-target edges applying the Data Processing Inequality (DPI) filter with a 0 tolerance. MRs analysis is performed by evaluating the overlap between each given regulon and the listed "hits" (Top DE genes) by two-side GSEA analysis.

In general, for each MR, the approach divides the MR's targets into positive (A) and negative (B) that were previously defined using Spearman's correlation, then plots on top, the DE ($\log_2$-FC of all genes) observed when comparing the experimental condition versus the control (this is called phenotype) in which genes are ranked from higher to lower $\log_2$-FC values. The observed differential enrichment scores (dES) is the difference of the GSEA statistics in the ranked phenotype of A minus B where large positive dES indicates an induced regulon status while a large negative dES indicates the opposite case (M. A. A. Castro et al., 2016; M. A. Castro, Wang, Fletcher, Meyer, & Markowetz, 2012; Groeneveld et al., 2019) (See Figure 3.9 C&D).

TME

Since the importance of the TME, the steps in the antitumor immune response, and the specific components and findings within the DLBCL or PCNSL context, has already been covered in Chapter 2.3, this section is rather focused on viewing the TME through the lens of RNA-seq analysis to unravel features such as the immunological composition of the infiltrate, immune expression signatures, and the immune repertoire (TCR/BCR repertoire).

Deciphering the TME composition as well as the relative contribution of each immune cell to it has been traditionally achieved, albeit expensive, by using IHC or flow cytometry. However, since transcriptome-wide sequencing data captured on bulk biopsies contain information from both the tumor and its infiltrates, sophisticated analyses can be applied to expression data to determine the presence or absence and relative abundance of CD8 T-cells and other immune cell types. Current RNA-seq methods for determining the TME composition use either a deconvolution approach (e.g. CIBERSORT, CIBERSORTX, TIMER, MCP-counter) or a GSA analysis (GSEA or GSVA) on curated gene lists (Becht et al., 2016; B. Li et al., 2016; Aaron M. Newman et al., 2015; Aaron M. Newman et al., 2019). The deconvolution approaches use a reference matrix composed of representative expression signatures for specific immune cells. The intuition is that the immune infiltrate is a mixture of different immune cell types that have distinct RNA expression profiles. If the RNA expression profiles of these immune cells are known, the RNA expression profile of the mixture can be modeled as a linear combination of the RNA expression profiles of the component cells (Lau, Bobe, & Khan, 2019). Nevertheless, their performance is highly affected by how the immune RNA profile was constructed since this can be cancer-specific. A practical example of this approach in the context of PCNSL is the TME study done by Marcelis et. al. covered in Chapter 2.3.4 (Marcelis et al., 2020).

On the contrary, GSA methods compute enrichment scores based on the ranked expression of curated gene lists which were previously associated with a specific cell type. Such enrichment scores allow for inter-sample comparison of the size of particular immune cell populations but are typically not directly interpretable as relative fractions of different cell types (Lau, Bobe, & Khan, 2019). Examples of this approach include the study of the DLBCL's TME by Kotlov et. al. (Chapter

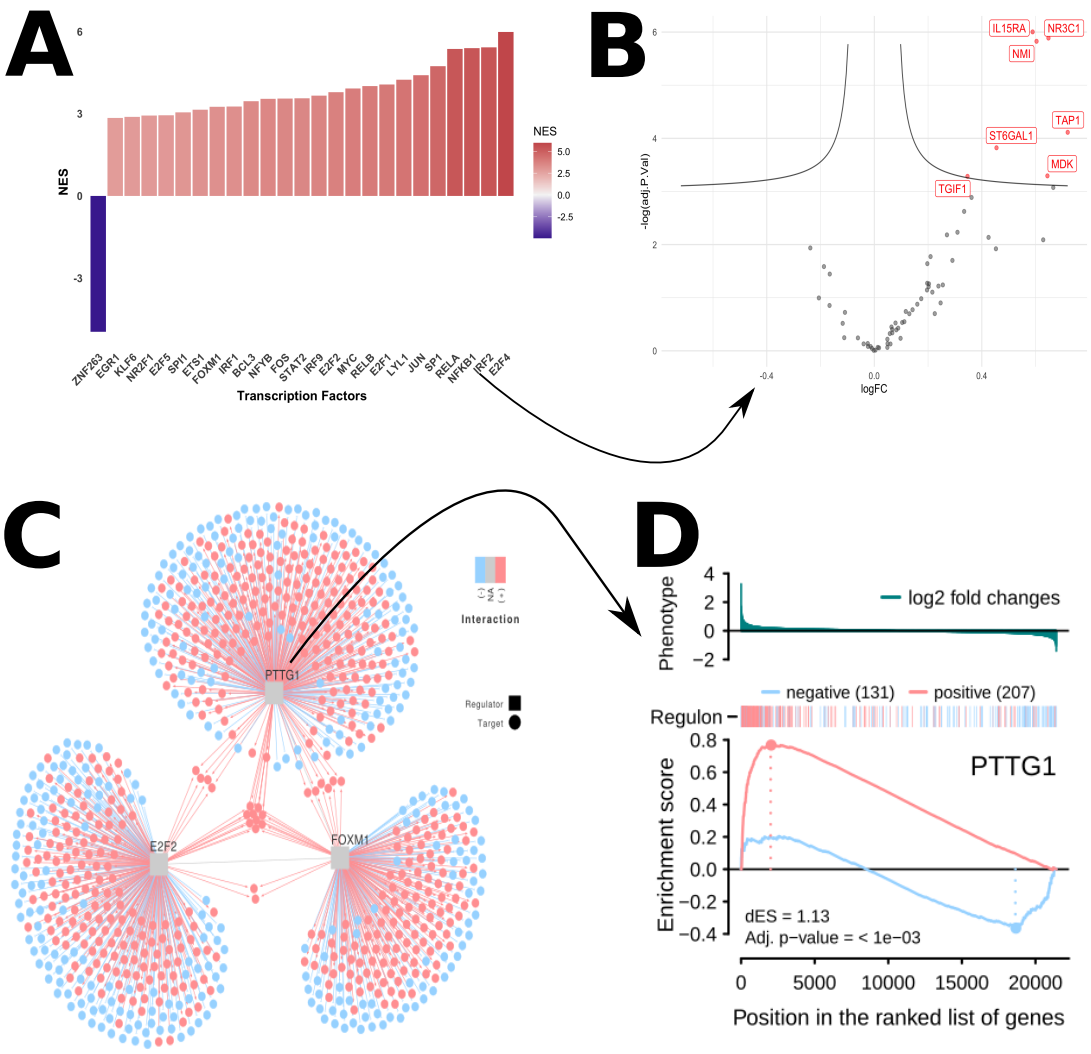


Figure 3.9: TFs and MRs activity calculation by DoRothEA and RTN. Example of DoRothEA's output showing the top 25 TFs NES (A), the expression of targets for IRF2 is later shown as a volcano plot (B). Example of the regulatory network of three TFs resulting from RTN-Reder packages (C) from which the MR activity of PTTG1 is later estimated by two-tailed GSEA (D). Panel D shows an overall induced regulon status for PTTG1 since most of its positive targets are upregulated. Adapted from Garcia-Alonso et. al. (2019) and Groeneveld et. al. (2019).

2.3.3) and another of PCNSL by Alame et. al. (Alame et al., 2021; Kotlov et al., 2021).

Immune expression signatures

Besides the relative abundance of immune cells in a tumor, the identification of specific phenotypic states of the immune microenvironment can also be achieved through the same GSVA/GSEA methods. Obtaining an immunological gene signature related to a specific phenotypic state is relatively simple; the idea behind it is identifying and characterizing DEGs between a condition of interest and a control condition. Examples of such signatures include the cytolytic index (CYT), the T-cell-inflamed phenotype, and the exhausted T-cell phenotype (Rooney, Shukla, Wu, Getz, & Hacohen, 2015; Schubert et al., 2018).

TCR/BCR features

Thanks to the capture of both expression and sequence information, RNA-seq can resolve the TCR and BCR immunological repertoires which have been associated with effective immunotherapy responses (Lau, Bobe, & Khan, 2019; Riaz et al., 2017; G. W. Wright et al., 2020). RNA-seq analysis on TCR can identify clonal expansion or reduction of T-cells which could be used as indicative of a productive or nonexistent antitumor immune response. On the other hand, BCR analysis can provide information about B-cell density, immunoglobulin rearrangements, and B-cell differentiation stages (Bolotin et al., 2015; Riaz et al., 2017; G. W. Wright et al., 2020). However, the TCR and BCR genes have been challenging regions to analyze using short-read NGS techniques due to their genetic diversity and variability. TCR/BCR receptors are encoded by three gene segments (VDJ, see Chapter 2.1.1), of which there are hundreds of alleles, that undergo recombination to generate a full-length receptor gene that is unique to that cell (the so-called clonotype) (Lau, Bobe, & Khan, 2019). Although being challenging, bioinformatic approaches to assembly TCR/BCR repertoires from short-read RNA-seq data, such as MiXCR and immunoseq have been reported (Bolotin et al., 2017; Morin et al., 2016). Typically, such methods involve aligning the reads to the TCR/BCR genes, performing further short-read assembly, finding identical sequences to group-specific clonotypes, correcting PCR/sequencing errors, and generating final clonotypes' counts (Figure 3.10) (Bolotin et al., 2017). Downstream analysis tools, such as immunarch and VDJtools, require the TCR/BCR clonotypes as input to generate information such as normalized unique clonotype counts, clonotype frequency distribution, rarefaction curves, clonotype tracking, repertoire overlap (between samples), and the total repertoire diversity which reflects the ability of our immune system to effectively withstand a multitude of encountered pathogens (Figure 3.11) (ImmunoMind Team, 2019; Shugay et al., 2015).

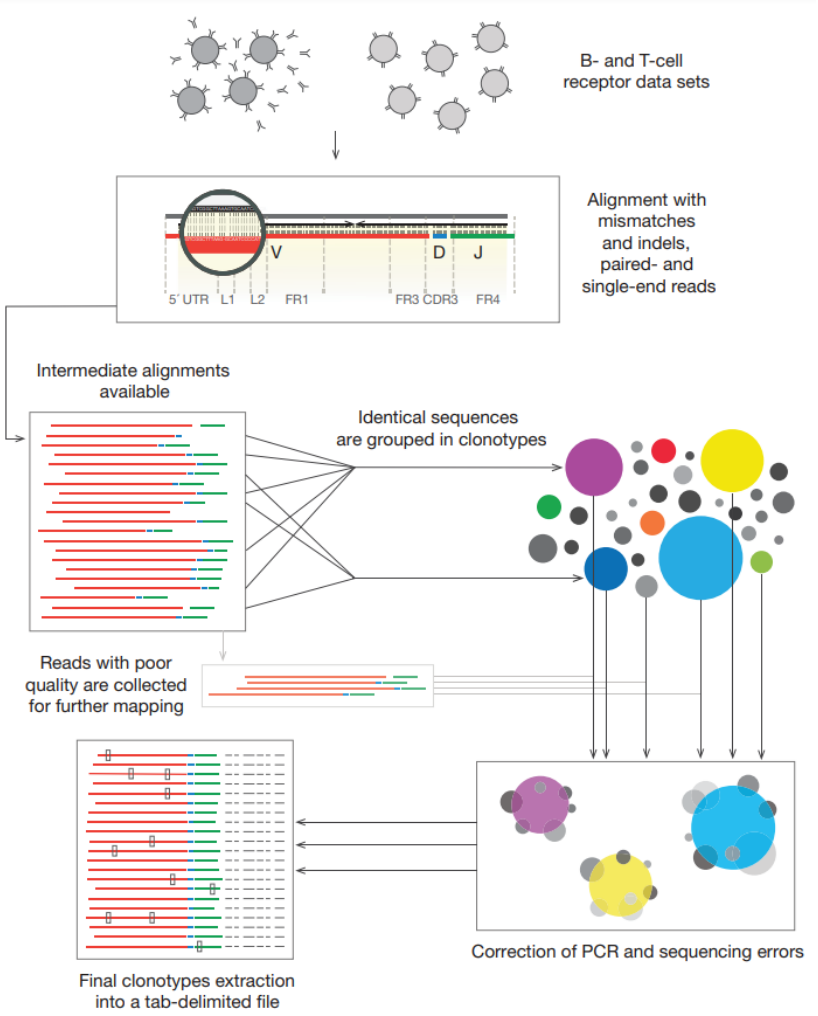


Figure 3.10: MiXCR pipeline. Example of a bioinformatic workflow for processing TCR/BCR data to extract clonotypes. Taken from Bolotin et. al. (2017).

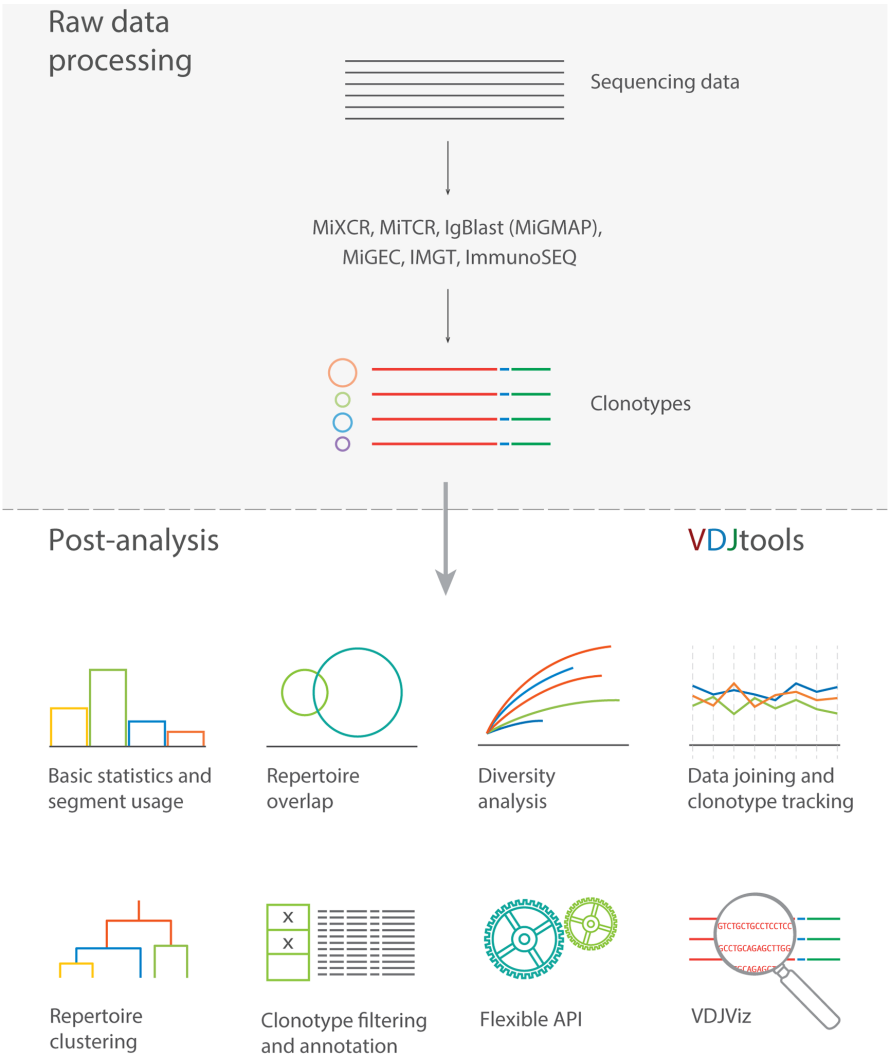


Figure 3.11: Overview of VDJtools downstream analyses. After generating clonotypes' reading, post-analysis options include general statistics (clonotype and read count, number and frequency of non-coding clonotypes, convergent recombination of CDR3 amino acid sequences, insert size statistics, etc), spectratyping (distribution of clonotype frequency by CDR3 length), Variable and Joining segment usage profiles, repertoire overlap analysis, among others. Taken from Shugay et. al. (2015).

3.1.4 ClinicOmic data

The ultimate objective of acquiring thousands of omic features is finding their clinical relevance, e.g. follow how patients respond to treatment over time given the presence of one or multiple omic features which is impossible in the absence of clinicOmic data. This term comprises the multiple features that are routinely evaluated in a clinical assessment of a patient such as the age, KPS, sex, diet, family history, treatments, OS, PFS, among others (Rennard, 2005). Given that the KPS and age definitions and implications in PCNSL were covered in Chapter 2.4.1, as well as the PCNSL treatments in Chapter 2.4, this sections aims to describe the most important concepts and methods in survival analysis for cancer research.

Time-to-event studies evaluate the long-term effects of new therapeutic regimens (or drugs) through a surrogate endpoint such as disease progression (i.e. PFS) or death (i.e. OS). A requirement of new drug approvals in oncology by the FDA and other regulatory bodies is showing direct/indirect clinical benefit when using either PFS or OS as endpoint, though the last one is universally more accepted (Hess, Brnabic, Mason, Lee, & Barker, 2019). Furthermore, time-to-event studies typically employ two closely related statistical approaches, KM and Cox proportional hazards (CoxPH) model analyses which are univariate and multivariate approaches, respectively (Dudley, Wickham, & Coombs, 2016).

KM analysis

KM, which is the most frequent survival analysis method used in randomized (phase III and some phase II) medical clinical trials, calculates how long after starting a particular treatment that the studied event (e.g. death, disease progression) occurred for individuals who were not otherwise censored. Patients from a sample become censored when investigators cannot determine if or when a subject ultimately experiences the event or otherwise drops out or is lost from the study) or at the end of the study (right censoring of all remaining subjects because no further data will be collected). Censoring is a major difference between KM and more traditional parametric analyses since missing data is a problem that can potentially bias data analysis and statistics (Figure 3.12 A&B).

The statistical output of a KM analysis can be graphed as a stair-step plot in which the x-axis represents the time variable expressed in a linear fashion (i.e. weeks, months, years, etc.) and the y-axis indicates the sample proportion that has not experienced the studied event. The length of each horizontal line represents the survival duration for that interval, and all survival estimates to a given point represent the cumulative probability of surviving to that time. The length of each interval is determined the appearance of an event which forms a downward step, whereas the tick marks indicate censored subjects. The probability of surviving an interval is related to the number of patients in that interval: Both the numerator and the denominator decrease by the number of patients who experienced the event plus those who were censored. Each of these probabilities

contributes to the subsequent and final probability of not experiencing the event. Moreover, the patients at risk at a specific time point is determined by the number of remaining patients that haven't experienced an event (Figure 3.12 C). Another advantage of KM curves is the rapid visualization of median survival which is the amount of time from the start of treatment that half of the patients from a cohort are still alive (Dudley, Wickham, & Coombs, 2016).

Statistically speaking, KM estimates are commonly calculated with the log-rank test in which chi-squares (χ^2) for each event time are summed to calculate an ultimate chi-square for each arm (or group). Log-rank results compare the full curves of each group and generate a significance level (p-value). For example, the KM curves in Figure 3.12D show a significant difference in outcome ($p < 0.01$) between cohort A (median survival = 11 months) and cohort B (median survival = 6.5 months). However, it is important to remark that the log-rank test allows between-group comparisons of survival estimates but not the size of a potential difference or of confounding variables such as age, sex or KPS (Dudley, Wickham, & Coombs, 2016). Programs to estimate survival by KM models include survminer, survival, and survtype (Terry M. Therneau & Patricia M. Grambsch, 2000).

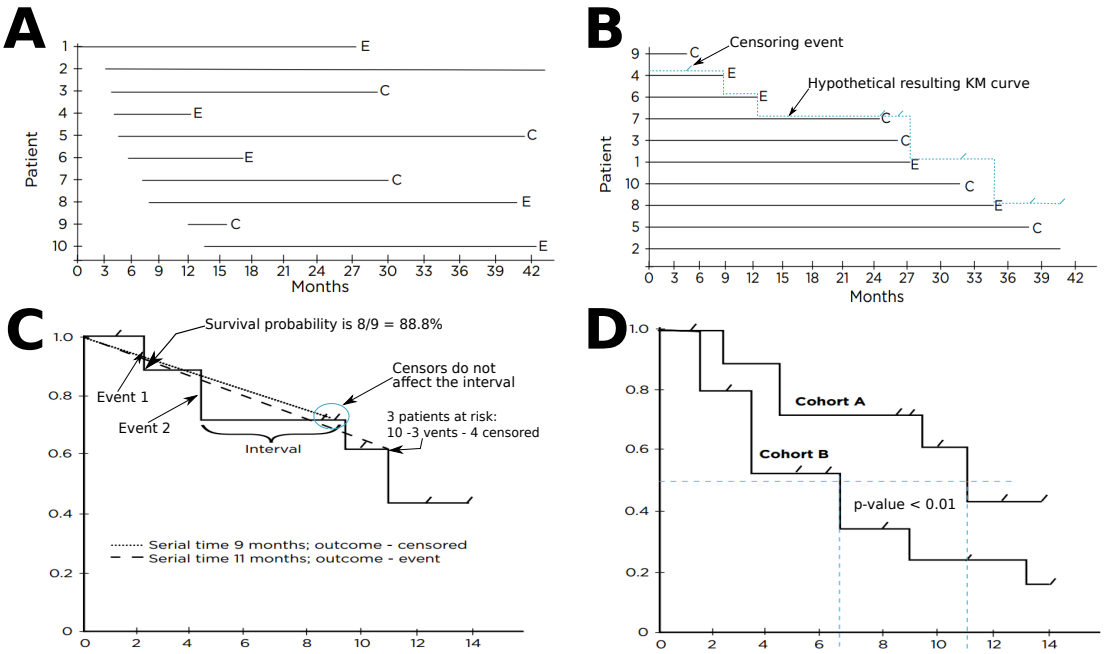


Figure 3.12: Interpreting a KM plot. Panel A shows each patient's serial time from a cohort (in sequential order) and whether they experienced the event (E) or were censored (C). Patient 2 has a continuing serial line since neither the endpoint for data collection nor an event has occurred. Panel B shows the same hypothetical patients arranged from the shortest to the longest serial time needed for plotting the KM graph (blue line). Panel C shows a hypothetical KM curve (cohort of 10 patients) along with some basic concepts such as survival probability and patients at risk. Panel D shows a significant difference in outcome ($p < 0.01$) where the median survival is 6.5 months in cohort B and 11 months in cohort A. Adapted from Dudley et. al. (2016).

CoxPH model analysis

As several confounding variables can potentially affect the interpretation of KM results, CoxPH is generally preferred since it allows adjusting survival with respect to several factors simultaneously while providing the effect size for each factor (hence the so-called multivariate survival model). The model permits examining how covariates influence the rate (so-called hazard rate) of the event happening at a particular point in time. Moreover, the Cox PH model uses only the rank ordering of the failure and censoring times and thus is less affected by outliers in the failure times than fully parametric methods. Briefly, the hazard function, $h(t)$, can be interpreted as the risk of dying at time t and can be estimated as follows:

$$h(t) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + ...\beta_n x_n) \tag{3.1}$$

where t is the survival time, $h(t)$ is the hazard function determined by a set of n covariates $(x_1, x_2, ...x_n)$, β_1, β_2 are the coefficients measuring the effect size of the covariates, and $h_0(t)$ is the baseline hazard function (for $x = 0$).

Interpreting the hazard ratio (HR) is counter-intuitive because a HR greater than one indicates that the value of the i^{th} covariate increases hence decreasing the length of survival. HRs are normally represented along with their upper and lower 95% confidence intervals which help determine if the observed effect is significant or not; if the 95% confidence interval (CI) "touches" a $HR = 1$, then the effect is considered not significant. The statistical significance by itself, calculated by Wald statistics, corresponds to the ratio of each regression coefficient to its standard error ($z = \beta_n / se(\beta_n)$) and evaluates whether the β_n coefficient of a given variable is statistically significantly different from 0 (Bender, Augustin, & Blettner, 2005). Programs for CoxPH modeling also include survminer, survival, and survtype (Terry M. Therneau & Patricia M. Grambsch, 2000).

Evaluating goodness of fit in survival models

Evaluating the performance of prognostic models is commonly achieved by using the concordance index (C-index) since it accounts for risk, outcome occurrence, and timing which enables distinguishing between well-behaved models and quasi-random ones. If we define M as the "risk score," i.e. the probability that an event might happen, T as some time point, and i/j as two generic subjects, then one can define the C-index as the following probability, conditioned on the relative order of events:

$$C = P(M_j > M_i | T_j < T_i) \tag{3.2}$$

This results in C-index values ranging from 0 to 1 where a "good" model, in the eyes of the C-index ($C = 1$), is one that always assigns higher scores to the subjects who experience the earlier events. Note that this may not always be the most appropriate definition of goodness of fit (e.g., when the highest risk is, in fact, related to long-term outcomes); nevertheless, it is the most common in survival analysis, where many techniques (KM, CoxPH, etc) assume the existence of a monotonic map between event probabilities and onset times (Longato, Vettoretti, & Di Camillo, 2020).

The de-facto standard way to compute the C-index is using Harrell's estimator ($\hat{\mathrm{C}}$) which works by estimating the ratio between the number of concordant and comparable pairs. A pair of subjects (i, j) is "comparable" if we can determine which of them (i or j) was the first to experience an event. A comparable pair is also "concordant" if the subject who experiences the earlier event is identified as the one having the greater risk, while "discordant" otherwise. It is defined as:

$$\hat{\mathrm{C}} = \frac{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} I(T_i < T_j) I(M_i > M_j)}{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} I(T_i < T_j)} \tag{3.3}$$

where $I(M_i > M_j)$ is the indicator function that can be equal to either 1 (results in $\hat{\mathrm{C}} = 1$), if all comparable pairs (i, j) have been assigned scores that reflect the correct order of events, or 0 (results in $\hat{\mathrm{C}} = 0$) if all pairs are discordant. Intermediate cases producing $\hat{\mathrm{C}} = 0.5$ denote a model that generates completely random assignments.

It is important to remark that interpreting intermediate values of the Harrell's C-index ($0.5 < \hat{C} > 1$) is not as straightforward as those resulting from binary classifications or even area under the ROC curve (AUROC), because the relationship between $\hat{C}$ and the proportion of subjects with incorrectly assigned risk scores (w_e) is not linear. For instance, while for AUROC an improvement from 0.75 to 0.80 means that a lower number of subjects (5% of the total) has been ranked incorrectly, for Harrell's C-index the interpretation depends in which area such increment falls. More precisely, the same change from 0.75 to 0.80 but in $\hat{C}$ decreases w_e by 8% (from $w_e = 0.71$ to $w_e = 0.63$); while improving $\hat{C}$ from 0.94 to 0.99 requires a 21% change ($\Delta w_e = 0.35 - 0.14$) (Longato, Vettoretti, & Di Camillo, 2020).

3.2 Reduction of multi-omic features

Each layer of omic data offers a large number of features which further increases when passing to the multi-omic step, hence making data manipulation a difficult and computationally demanding task. Finding the most effective features among thousands of other ones in a feature selection/reduction process is a fundamental challenge in the field of omics data analysis. Common techniques for data reduction can be classified as feature reduction (i.e. select a subset of features with the least redundancy and the most relevance to the target class in order to obtain the highest classification accuracy) and feature extraction (i.e. produce new features with lower dimensions than the main features). The feature reduction method provides a better understanding of the system by selecting the best not redundant feature sets at a low computational cost (Momeni, Hassanzadeh, Saniee Abadeh, & Bellazzi, 2020). Given that feature reduction methods are often preferred, this section is focused on overviewing the main approaches (filter, wrapper, embedded, and ensemble; see Figure 3.13).

Filter approach

This approach is particularly useful for huge datasets, it measures the characteristics of features based on four types of evaluation criteria: Dependency, Information, Distance, and Compatibility. Filtering methods can be further divided into uni-variate which uses the standard deviation (sd), the median absolute deviation (mad), the mutational frequency or specific manual cut-off based on the literature, and multivariate methods which aim to find relationships between features. An example of multivariate methods is combining clinicomic data (outcome, sex, age) with gene expression data to evaluate which genes are associated with survival when performing a multivariate or univariate CoxPH analysis (Lu, Meng, Zhou, Jiang, & Yan, 2021; Momeni, Hassanzadeh, Saniee Abadeh, & Bellazzi, 2020). Other methods first reduce the dimensions of feature space and then apply filter methods such as the principal component analysis (PCA).

Wrapper approach

Wrapper approaches, which demand a higher computational cost compared to filter approaches, select a subset of discriminating features by minimizing the

prediction error of a particular classifier. Because the number of subsets of features expands exponentially as a function of the initial input, this approach can become highly demanding when having thousands of features. Moreover, the approach has the potential to overfit on data with a small number of samples (Momeni, Hassanzadeh, Saniee Abadeh, & Bellazzi, 2020).

Embedded approach

Based on learning algorithms, the embedded approach results more efficiently and computationally less complicated than the wrapper approach because it avoids duplicate execution and examines each feature subset in the learning process. The most popular embedded method is support vector machine-recursive feature elimination (SVM-RFE) which uses an iterative backward selection approach to remove least-weighted features between two "classes" during each iteration; however, it depends on the manual labeling of each "class" (Momeni, Hassanzadeh, Saniee Abadeh, & Bellazzi, 2020).

Ensemble approach

The idea behind this approach is to reduce the dataset into a number of smaller subsets and then rely on different feature selection strategies for each subset to produce a merged result from these groups. Sub-setting allows less tendency to overfitting, more stability, and less dependency on the algorithms (Momeni, Hassanzadeh, Saniee Abadeh, & Bellazzi, 2020).

3.3 Multi-omic data integration

Despite combining several different omics data to discover coherent biological signatures being considered the most challenging step, it is inevitable to mention that it relies on how well data acquisition and reduction were conducted. The objective of multi-omic data integration is to incorporate the different biological single-omic layers of information to reconstruct the complex interdependent interactions shared within a molecular subtype of the disease and to predict phenotypic outcomes (tumor/normal, early/late stage, survival, etc.) (Huang, Chaudhary, & Garmire, 2017; Lu, Meng, Zhou, Jiang, & Yan, 2021). The clustering algorithms to be presented in this section have the ultimate goal of finding such molecular subtypes, however, it is important to note that finding the "correct" method/result requires a profound biological and bioinformatic background of the disease for achieving an integrated interpretation of the findings.

3.3.1 Clustering methods

The first and most important parameter when applying any clustering algorithm is estimating the optimum number of clusters k for the data, where k needs to be small enough to reduce noise but large enough to retain important information (Lu, Meng, Zhou, Jiang, & Yan, 2021). The most used estimators of k are the Gap-statistics and the Cluster Prediction Index (CPI) (Chalise & Fridley, 2017; Tibshirani, Walther, & Hastie, 2001). The idea behind Gap-statistics is finding

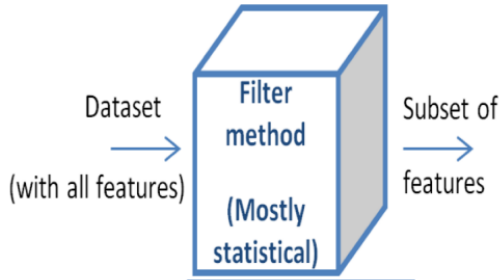
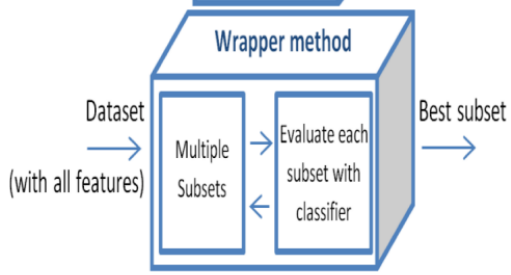
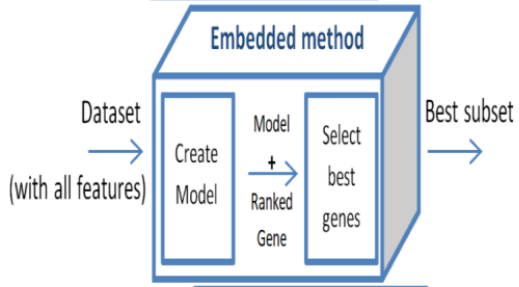
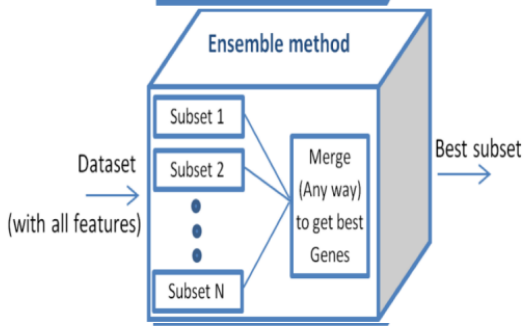
Method	Shape	Advantage	Disadvantage
Filter		Fast and scalable Independent of classifier Faster than wrapper methods Better computational complexity than wrapper methods	Less accuracy due to non-consideration of the classifier Ignores the relationship between features / variables. May involve redundancy.
Wrapper		Interact with classifier Interaction between features Better accuracy than the filter methods	Overfitting High computational complexity Expensive operation Tends to local optimization
Embedded		Higher Accuracy and efficiency than filter methods Less computational complexity than wrapper methods More focus on relationship between features	Dependent to classifier
Ensemble		Less tendency to overfitting More scalability for high-dimensional data sets Stability	Understanding the combination of classifiers is difficult

Figure 3.13: Comparison between feature reduction methods. Adapted from Momeni et. al. (2020).

a k value which maximizes the $Gap_n(k)$ estimate by comparing the change in within-cluster dispersion with that expected under an appropriate reference null distribution. $Gap_n(k)$ is defined as:

$$Gap_n(k) = E_n^*\{log(W_k)\} - log(W_k) \tag{3.4}$$

where E_n^* denotes expectation under a sample of size n from the reference distribution and W_k is the error measure (within-cluster dispersion) which is affected by n data points under p dimensions (Tibshirani, Walther, & Hastie, 2001).

A practical example of its interpretation (in the genomic area) is given by Tibshirani et. al. where they applied hierarchical clustering to DNA microarray data coming from nine different tumor types (Figure 3.14 A). Their gap-statistic methodology estimated the best k value to be 2, 6, or 8 as the $Gap(k)$ function rises when finding both values (Figure 3.14 B). However, the authors remark that it is important to examine the entire gap curve rather than only taking the highest position (Tibshirani, Walther, & Hastie, 2001).

The CPI is the average of adjusted rand indices which are calculated iteratively by comparing clustering assignments between the "observed" clusters (from training data) and the "predicted" clusters (from the test data). CPI values range from 0 to 1 where a higher value indicates a good consensus between the predicted and the observed clustering assignments (Figure 3.14 C). The integrative non-negative matrix factorization (intNMF), which is an extension of the NMF algorithm and is the behind CPI calculation, is integrated into an R package named "intNMF" (Chalise & Fridley, 2017). After finding the optimum number of clusters, the next step is properly applying either one or a combination of clustering methods such as NMF, iCluster, iCluster+, ConsensusClustering, and Perturbation clustering for data INtegration and disease Subtyping (PINSPlus) (Hoadley et al., 2014; Huang, Chaudhary, & Garmire, 2017; Lu, Meng, Zhou, Jiang, & Yan, 2021; Monti, 2003; Nguyen, Shrestha, Draghici, & Nguyen, 2019).

NMF

This method is based on decomposing a non-negative matrix into non-negative loadings (coefficients) and non-negative factors:

$$min||X - WH||^2, W \geq 0, H \geq 0 \tag{3.5}$$

where X is the matrix a layer of omic data that has $M \times N$ dimensions, W is the common factor for $M \times K$ dimension matrix and H is the $K \times N$ dimension coefficient matrix. Of note, this method requires previous data normalization of each layer and/feature (Huang, Chaudhary, & Garmire, 2017).

iCluster and iCluster+

Unlike NMF, these methods do not require non-negative input data and add, besides the W and H, an E value to represent the error/noise term. Additionally, the iCluster+ assumes different modeling assumptions for each omic layer (e.g. logistic, normal linear, multilogit, and Poisson distributions) and integrates least absolute shrinkage and selection operator (LASSO) to address the sparsity issue in H (Huang, Chaudhary, & Garmire, 2017).

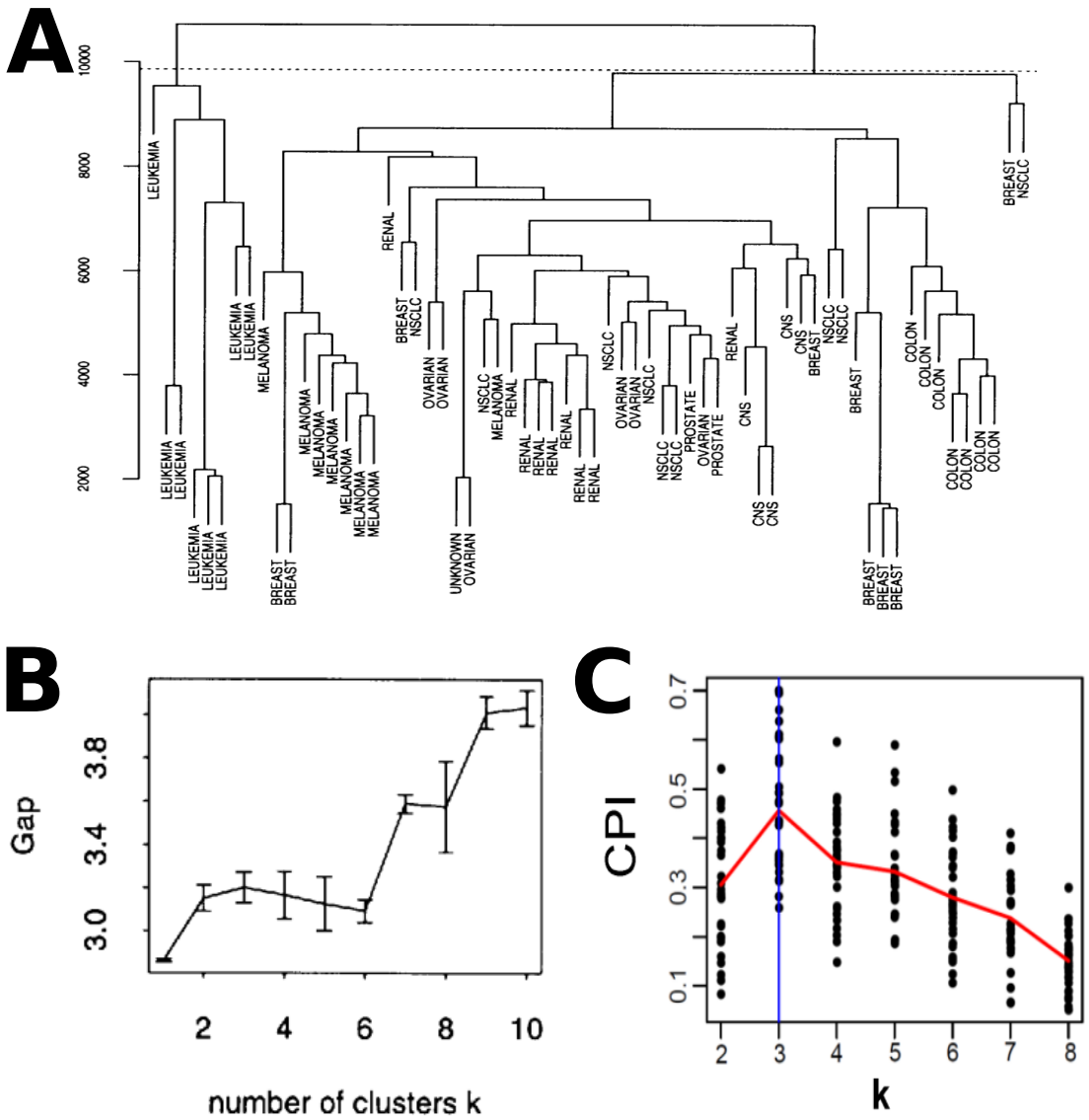


Figure 3.14: Gap-statistics and CPI methods for finding optimal cluster number. Dendrogram from hierarchical clustering of DNA microarray data coming from nine different tumor types (Panel A) and gap-statistic as a function of the number of clusters (Panel B). CPI plot resulting from running the intNMF algorithm on glioblastoma TCGA multi-omic data (CNA, methylation, and mRNA). Adapted from Tibshirani et. al. (2001) and Chalise and Fridley (2017).

ConsensusClustering

The idea behind the ConsensusClustering is to represent the consensus over multiple runs of a clustering algorithm with random restart (such as K-means, model-based Bayesian clustering, self-organizing map (SOM), among others), to account for its sensitivity to the initial conditions (Hoadley et al., 2014; Monti, 2003).

PINSPlus

Nguyen et. al. developed an unsupervised two-step approach for subtype discovery without using any a priori knowledge. When integrating multi-omics data, the first step aims to identify subgroups that are strongly connected across all data layers by merging connectivities (i.e. features) of all data layers into a similarity matrix that represents the overall connectivity between patients. Step two splits each discovered group individually to avoid over-splitting in the first step but only if either stage I clustering was extremely imbalanced or there are not strong signals across all data layers within a subtype (Nguyen, Shrestha, Draghici, & Nguyen, 2019).

Cluster Ensembles

Similar to ConsensusClustering, this method's goal is to improve clustering by finding consensus results from different clustering methodologies. Cluster Ensembles selects the best solution for a given situation from three different techniques including reclustering based on similarity measures from the partitionings (I), from hypergraph partitionings (II), and from collapsing meta-clusters (Strehl & Ghosh, 2002).

Chapter 4

Thesis objectives

This thesis aims to characterize the multi-omic landscape of PCNSL, including genomics, epigenomics, transcriptomics, and clinicomics, and to integrate such data to find molecular PCNSL subgroups with biological and clinical relevance. Ergo, the objectives are divided into three parts:

- **Chapter 4.1:** To review the literature to understand the HLA structure/diversity and genetic susceptibility in PCNSL and other B-cell NHLs.

- **Chapter 4.2:** To develop a code to track c-AID mutations and to explore their implications at pan-cancer level (~ 50,000 samples).

- **Chapter 4.3:** To extract, analyze, and integrate multi-omic data to find and characterize molecular PCNSL subgroups with shared causative biologic factors of disease and outcome.

4.1 Review of the literature to understand the HLA structure/diversity and genetic susceptibility in PCNSL and other B-cell NHLs

B-cell NHLs' main subtypes include DLBCL, FL, CLL and extra-nodal lymphomas (i.e. brain, eyes, leptomeninges, spinal cord, etc.). Their risk associations had been mainly attributed to family history of the disease, inflammation, and immune components including HLA genetic variations. Nevertheless, a broad range of genome-wide association studies (GWAS) have shed light into the identification of several genetic variants presumptively associated with B-cell NHL etiologies, survival or shared genetic risk with other diseases.

In this review, I used published articles to overview the HLA structure and diversity and summarize the evidence of genetic variations, by GWAS, on five NHL subtypes (diffuse large B-cell lymphoma DLBCL, follicular lymphoma FL, chronic lymphocytic leukemia CLL, marginal zone lymphoma MZL, and primary central nervous system lymphoma PCNSL).

The review article was published at the International Journal of Molecular Sciences (10.3390/ijms22010122) and is listed in the Results section.

4.2 Exploring the implications of AID-related mutations at pan-cancer level

Activation-induced cytidine deaminase, AID (encoded by *AICDA*), is a driver of somatic hypermutation and class-switch recombination in immunoglobulin related genes within naive B-cells. This AID deamination of cytosine to uracil also occurs during IG gene transcription and inside particular DNA patterns such as WA motifs (W = A/T) or WRCY motifs (R = purine; Y = pyrimidine). Mutations arising at WA motifs are defined as non-canonical AID (COSMIC signature 9), whereas those arising at WRCY motifs as canonical-AID. Furthermore, within hematological cancers off-target AID activity has been reported responsible of lymphomagenesis, maninly through the mutations activating NF-$\kappa\beta$ pathway. In addition, this deaminase belonging to the APOBEC family may have off-target effects genome-wide, but despite this, no detailed characterization of the involvement of AID-related mutations at the pan-cancer level, as well as their potential mutational and clinical implications, has been performed.

Here, I used more than 50.000 samples covering more than 80 tumor types at the bulk level and close to 2.5 million cells at single-cell resolution to thoroughly describe the landscape of AID-related mutations. Furthermore, the developed and validated code to track the c-AID mutations served to fully characterize such mutations in PCNSL for the following section.

The original research article is currently under revision in the Cell Reports journal (Pre-print available) and is listed in the Results section.

4.3 Integrating multi-omic data to characterize PCNSL molecular and clinical diversity

PCNSL is a rare subtype of extranodal non-Hodgkin's lymphoma, in the vast majority of cases consist of diffuse DLBCLs histologically, but has a less favorable prognosis (median OS of 26 months versus 124 months) and has been proved to be molecularly a different biological entity (Chapuy et al., 2018; Schmitz et al., 2018; Sehn & Salles, 2021; Yoshida et al., 2016). The gold standard treatment, high-dose methotrexate regimen, is often associated with neurotoxicity and eventual relapse (up to 60% of the patients) (Houillier et al., 2020; Y. Zhou et al., 2018). Moreover, recurrent/relapsed PCNSLs have shown heterogeneous responses in diverse clinical trials with different treatment strategies (Garcilazo-Reyes et al., 2020).

In addition, PCNSL was initially found to be at late B-cell germinal center stages and to have constitutive NF-$\kappa\beta$ activity due to mutations in genes of the BCR pathway (*CD79B*, *SHIP*, *CBL*, and *BLNK*), of the TLR pathway (*MYD88*) and *CARD11*; however, recent studies have pinpointed it to belong to the so-called "MCD" or Cluster 5 (C5) DLBCL which converge in the presence of frequent *MYD88* (L265P), *CD79B*, *PIM1*, *BTG2* mutations, IgH-BCL6 translocations, copy gains of 3q12.3, 9p24.1 (PD-L1/PD-L2), 11q and copy losses of 6p21-22 (HLA locus), 6q21, and 9p21.3 (*CDKN2A* biallelic loss). Interestingly, the AID has a higher off-target mutagenic activity in PCNSL compared to other DLBCL (Chapuy et al., 2018; Schmitz et al., 2018; G. W. Wright et al., 2020).

Although DLBCL classification based on genomics has improved clinical decision making, PCNSL heterogeneity has not been properly addressed mainly due to the lack of multi-omic data integration and the limited number of patients.

Here, I extracted and integrated distinct multi-omic features, such as mutations, copy-number alterations, fusions, gene expression, TCR/BCR clonotypes, TME, methylation, radiological characteristics, OS times, and PFS times, from a total of 147 immunocompetent, treatment naïve PCNSL patients. This data allowed me to find and thoroughly characterize PCNSL molecular subtypes which I validated in a second FFPE cohort of 93 patients. Additionally, to facilitate routine clinical implementation, I developed an algorithm (RBraLymP) that uses gene expression data from either FFPE or FF tissue, to identify the PCNSL molecular subtypes associated with multi-omic features.

The original research article is currently under revision at Nature Medicine (Pre-print available) and is listed in the Results section.

Chapter 5

Results

- **Manuscript 1 (Published): HERNANDEZ-VERDIN Isaias**, LABRECHE Karim, BENAZRA Marion, MOKHTARI Karima, HOANG-XUAN Khê, ALENTORN Agusti (2020). Tracking the Genetic Susceptibility Background of B-Cell Non-Hodgkin's Lymphomas from Genome-Wide Association Studies. International Journal of Molecular Sciences.
- **Manuscript 2 (Under review): HERNANDEZ-VERDIN Isaias**, AKDEMIR Kadir, RAMAZZOTTI Daniele, CARAVAGNA Giulio, LABRECHE Karim, MOKHTARI Karima, HOANG-XUAN Khê, PEYRE Matthieu, BIELLE Franck, TOUAT Mehdi, IDBAIH Ahmed, DUVAL Alex, SANSON Marc, ALENTORN Agusti (2022). Pan-cancer landscape of AID-related mutations, composite mutations and their potential role in the ICI response. Cell Reports.
- **Patent 1 (Submitted): HERNANDEZ Isaias**, ALENTORN Agusti (2021). ACTIVATION-INDUCED CYTIDINE DEAMINASE AS A NEW BIOMARKER (European Patent Application No. EP21305876.1). European Patent Office. Available in the APPENDIX 2.
- **Manuscript 3 (Under review): HERNANDEZ-VERDIN Isaias**, KIRASIC Eva, WIENAND Kirsty, MOKHTARI Karima, EIMER Sandrine, LOISEAU Hugues, ROUSSEAU Audrey, PAILLASSA Jérôme, AHLE Guido, LERINTIU Felix, URO-COSTE Emmanuelle, OBERIC Lucie, FIGARELLA-BRANGER Dominique, CHINOT Olivier, GAUCHOTTE Guillaume, TAILLANDIER Luc, MAROLLEAU Jean-Pierre, POLIVKA Marc, ADAM Clovis, URSU Renata, SCHMITT Anna, BARILLOT Noemie, NICHELLI Lucia, LOZANO-SANCHEZ Fernando, IBAÑEZ-JULIÁ Maria-José, PEYRE Matthieu, MATHON Bertrand, ABADA Yah-se, CHARLOTTE Frédéric, DAVI Frédéric, STEWART Chip, DE REYNIÈS Aurélien, CHOQUET Sylvain, SOUSSAIN Carole, HOUILLIER Caroline, CHAPUY Bjoern, HOANG-XUAN Khê, ALENTORN Agusti (2022). Molecular and clinical diversity in primary central nervous system lymphoma. Nature Medicine.

5.1 Tracking the Genetic Susceptibility Background of B-Cell Non-Hodgkin's Lymphomas from Genome-Wide Association Studies

Isaias Hernández-Verdin,[1,2] Karim Labreche,[1,2] Marion Benazra,[1,2,3,4,5] Karima Mokhtari,[6,7] Khê Hoang-Xuan,[1,2,3,4,7,8] and Agusti Alentorn[1,2,3,4,7,*]

[1]Faculté de Médecine, Sorbonne Université, 75013 Paris, France; isaias.hernandez@icm-institute.org (I.H.-V.); karim.labreche@icm-institute.org (K.L.); marion.benazra@icm-institute.org (M.B.); khe.hoang-xuan@aphp.fr (K.H.-X.)

[2]Brain and Spine Institute (ICM), 75013 Paris, France

[3]National Institute of Health and Medical Research (Inserm) U 1127, 75013 Paris, France

[4]National Center for Scientific Research, Joint Research Unit 7225, 75013 Paris, France

[5]Brain and Spine Institute (ICM), iGenSeq Platform, 75013 Paris, France

[6]Raymond Escourolle Department of Neuropathology, Public Assistance–Hospitals of Paris, Hospital Group of Pitié-Salpêtrière, 75013 Paris, France; karima.mokhtari@aphp.fr

[7]Assistance Publique Hôpitaux de Paris (APHP), Department of Neurology-2, Groupe Hospitalier Pitié Salpêtrière, 75013 Paris, France

[8]Réseau Expert National LOC (Lymphomes Oculo-Cérébraux), Groupe Hospitalier Pitié Salpêtrière, 75013 Paris, France

[*]Correspondence: agusti.alentorn@aphp.fr

Abstract

B-cell non-Hodgkin's lymphoma (NHL) risk associations had been mainly attributed to family history of the disease, inflammation, and immune components including human leukocyte antigen (HLA) genetic variations. Nevertheless, a broad range of GWAS have shed light into the identification of several genetic variants presumptively associated with B-cell NHL etiologies, survival or shared genetic risk with other diseases. The present review aims to overview HLA structure and diversity and summarize the evidence of genetic variations, by GWAS, on five NHL subtypes (diffuse large B-cell lymphoma DLBCL, follicular lymphoma FL, chronic lymphocytic leukemia CLL, marginal zone lymphoma MZL, and primary central nervous system lymphoma PCNSL). Evidence indicates that the HLA zygosity status in B-cell NHL might promote immune escape and that genome-wide significance variants can give biological insight but also potential therapeutic markers such as WEE1 in DLBCL. However, additional studies are needed, especially for non-DLBCL, to replicate the associations found to date.

Keywords: B-cell non-Hodgkin's lymphoma, GWAS, cancer risk, HLA

5.1.1 Introduction

Malignant lymphomas are among the most common head and neck neoplasm from lymphoreticular system origin and can be defined as Hodgkin's or non-Hodgkin's lymphoma (NHL), in which approximately 25% arises from extra-nodal locations like Waldeyer's ring, oral cavity, salivary glands, thyroid, larynx, nasal cavity, paranasal sinuses, skin, brain, eyes, leptomeninges, or spinal cord (Bowzyk Al-Naeeb, Ajithkumar, Behan, & Hodson, 2018; Hoang-Xuan et al., 2015). According to the 2016 World Health Organization classification there are around 60 distinct subtypes of NHL with diffuse large B-cell lymphoma (DLBCL, about 30%), follicular lymphoma (FL, about 20%) and chronic lymphocytic leukemia/small lymphocytic leukemia (CLL/SLL) among the most common (Ekström-Smedby, 2006; Singh et al., 2020; Steven H. Swerdlow et al., 2016). Along with the wide location distribution of the lymphomas, this group of diseases has varying etiologies and prognosis, for example the five-year survival is 85% for CLL, 80% for FL, 76.5% for marginal zone lymphoma (MZL), but <50% for more aggressive lymphomas such as DLBCL or even 30% for primary central nervous system lymphoma (PCNSL) (Braggio et al., 2015; Olszewski & Castillo, 2013; Zhong, Cozen, Bolanos, Song, & Wang, 2019). Furthermore, there have been vast population studies to associate the presence of different etiologies, such as smoking, height, weight, autoimmune conditions, alcohol consumption, viral infections, and genetics, with the risk of developing any subtype of NHL (Ekström-Smedby, 2006; Moore et al., 2020; Morton et al., 2014). Nevertheless, despite these efforts, there are only few established risk factors including autoimmune conditions (e.g., Sjögren disease, rheumatoid arthritis, systemic lupus erythematosus, and multiple sclerosis), immunodeficiency syndromes, organ transplants, breast implants and specific infections (e.g., *Helicobacter pylori* for mucosa-associated lymphoid tissue lymphoma of the stomach, immunodeficiency virus, and mononucleosis) (Din et al., 2019; Hjalgrim et al., 2010; K. E. Smedby & Ponzoni, 2017).

More recently, sequencing technologies like next generation sequencing (NGS) and genome-wide association studies (GWAS), have broaden the possible candidates by using thousands of genetic variants for multiple genetic risk factors identification (Mills & Rahal, 2019). GWAS combining the population structure (Q) jointly with the genetic marker based kinship matrix (K) mixed linear model, also called linear mixed model, where the test statistic for significance is drawn from the central Chi-square distribution by comparing the allele frequencies of the cases to the controls. A variant is said to be significant at genome-wide level if the p value is $\leq 5 \times 10^{-8}$, which was set by taking a 0.05 significance level and roughly dividing by the total number of independent blocks of linked genes in Europeans (thought to be 1,000,000) (M. Wang & Xu, 2019; Zhong, Cozen, Bolanos, Song, & Wang, 2019). Regarding GWAS within the B-cell NHL context, most studies have focused on genetic variants at chromosome 6p21, specifically human leukocyte antigen (HLA) variants, since that region is critical for innate and adaptive immune responses, but there have been also efforts to find associations with variants outside this chromosome and other etiologies (Bernatsky et al., 2017; Di Paolo et

al., 2019; Din et al., 2019; Mathilde R. W. de Jong et al., 2018; Kleinstern, Camp, et al., 2020; Moore et al., 2020; Zhong, Cozen, Bolanos, Song, & Wang, 2019).

In this review, we give an overview of HLA structure and diversity, and then we summarize the most recent GWAS presented in five B-cell NHL subtypes: DLBCL, FL, CLL, MZL, and PCNSL in the light of loci (HLA and others) diversity and zygosity specific associations in addition to further clinical evaluations when suitable. Furthermore, we present the genetic overlap between B-cell NHL subtypes and autoimmune diseases, height, lipid traits, or other lymphomas.

5.1.2 HLA Overview

The association found between Hodgkin's lymphoma (HL) and HLA-B gene variation allowed to the later discovery that the major histocompatibility complex (MHC) is the genomic region with the highest number of associated human diseases (Amiel, 1967; Trowsdale & Knight, 2013). The MHC, a hyper gene-dense region located at chromosome 6p21.3, encodes a set of 21 protein-coding loci that gives rise to three different types of HLA molecules. Firstly, class I, encoded by the highly polymorphic *HLA-A*, *HLA-B*, and *HLA-C* genes ("classical"); but also *HLA-E*, *HLA-F*, and *HLA-G* genes ("nonclassical") both of which are composed of a single α chain non-covalently bound to a small β2-microglobulin polypeptide encoded by another chromosome (15q21). Secondly, class II, encoded by the *HLA-DPA1*, *HLA-DQA1*, *HLA-DQA2*, *HLA-DRA*, *HLA-DPB1*, *HLA-DQB1*, *HLA-DQB2*, *HLA-DRB1*, *HLA-DRB2*, *HLA-DRB3*, *HLA-DRB4*, and *HLA-DRB5* genes which are composed of α-β heterodimer. Thirdly, class III, encoded by 61 genes (ex. *MIC*, *SKI2W*) involved in inflammation, leukocyte maturation and complement cascade (Deakin et al., 2006; Dendrou, Petersen, Rossjohn, & Fugger, 2018; Sanchez-Mazas, 2020; Zhong, Cozen, Bolanos, Song, & Wang, 2019). These HLA genes comprise approximately four million base pair region, giving rise to more than 15,000 different classical HLA class I and II alleles which can, theoretically, serve for presenting over 10^{12} different peptides if the antigen-presenting cell (APC) is heterozygous at each of the six classical class I or II HLA loci (J. Robinson, Soormally, Hayhurst, & Marsh, 2016; Sewell, 2012).

Nucleated cells express HLA class I molecules where small peptides (8–10 amino acids) are bound to the α1-α2 domains at the HLA peptide-binding site for later recognition by $\alpha\beta$ T-cell receptors (TCRs) on CD8+ T cells. On the other hand, monocytes/macrophages, dendritic cells and B cells express HLA class II receptors which can present larger peptides (13–25 amino acids) to TCRs on CD4+ T cells, hence inducing the orchestrated immune response against the pathogen due to cytokines releasing either by helping B cells to secrete high affinity antibodies or by inducing macrophage activation (Dendrou, Petersen, Rossjohn, & Fugger, 2018; Sanchez-Mazas, 2020; Zhong, Cozen, Bolanos, Song, & Wang, 2019). Furthermore, HLA class I molecules can bind to natural killer (NK) cells through immunoglobulin-like receptors and C-type lectin-like CD94/NKG2 receptors (Figure 5.1) (Saunders et al., 2015).

Along with the extreme gene density and polymorphism at the MHC locus,

linkage disequilibrium, low-throughput methodologies and samples sizes made HLA-disease associations complicated. However, with the advent of both high-throughput whole-genome-based methodologies (example GWAS) and the evolution of big data analysis, researchers can measure the contribution of a single genetic variation across the genome on a disease risk by leveraging linkage disequilibrium (M. Wang & Xu, 2019).

5.1.3 GWAS in B-Cell NHL

DLBCL

Representing around 30% of NHL and affecting preferentially older adults, diffuse Large B-cell lymphoma is the most common type of NHL. During the last two decades, treatment with immunochemotherapy consisting of cyclophosphamide, doxorubicin, vincristine, and prednisolone combined with the anti-CD20 monoclonal antibody rituximab (R-CHOP) has become the gold standard. This regimen results in a cure rate of 60% and a five-year survival rate of 60% for germinal center B-cell (GCB) subtype or 35% for activated B-cell (ABC) subtype; however, clinical course is heterogeneous, even after elucidation of cell of origin (ABC or GCB), thus requiring new biomarkers for elucidating patient outcome and for adapting treatment strategies (Ghesquieres et al., 2015; G. Wright et al., 2003). Different risks factors have been identified. A stepwise logistic regression meta-analysis of 4667 cases and 22,639 controls, found that DLBCL is associated with B-cell activating autoimmune diseases (odds ratio [OR] = 2.36, 95% confidence interval [CI] = 1.80 to 3.09), hepatitis C virus seropositivity (OR = 2.02, 95% CI = 1.47 to 2.76), family history of NHL (OR = 1.95, 95% CI = 1.54 to 2.47), and higher young adult body mass index (OR = 1.58, 95% CI = 1.12 to 2.23, for 35+ vs 18.5 to 22.4 kg/m^2). Conversely, different potential presumptive protective factors have been proposed, such as higher sun exposure (OR = 0.78 and 0.80, 95% CI = 0.69 to 0.89 and 0.71 to 0.90), in two studies, and lifetime alcohol consumption (OR = 0.57, 95% CI = 0.44 to 0.75, for >400 g vs nondrinker) in one study. Vitamin D deficiency has been suggested as negative prognostic factor in patients with aggressive DLBCLs but was not found to be associated with dietary intake (OR = 1.03, 95% CI = 0.90 to 1.19), hence indicating that other factors rather than vitamin D may be involved (J. R. Cerhan et al., 2014; Park, Hong, Lee, & Koh, 2019). More recent studies evaluating DLBCL risk, one using lipid trait variants in 2661 cases and 6221 controls found positive association with high density lipoproteins (OR = 1.14; 95% CI, 1.00–1.30), while another study evaluating height as variable; however, neither of them were significant after adjusting for multiple testing (Kleinstern, Camp, et al., 2020; Moore et al., 2020).

HLA-B (rs2523607) locus has been associated with DLBCL risk, initially described in a GWAS study with 3857 cases/7766 controls from European population (OR = 1.32; 95% CI = 1.21–1.44; $p = 2.40 \times 10^{-10}$) and then reported in 1124 patients and 3596 controls from Asian population (OR = 3.05; 95% CI = 1.32–7.05; $p = 9.0 \times 10^{-3}$), though not reaching genome-wide significance (Bassig et al.,

2015; James R. Cerhan et al., 2014) (Table 5.1). Regarding pleiotropy of DLBCL with other diseases, a European population study found that variant rs10484561 (*HLA-DQB1*01:01 DQA1*01:01 DQB1*05:01* extended haplotype, Linkage Disequilibrium (LD) $r^2 = 1.0$) is associated with both DLBCL risk (OR = 1.36; 95% CI = 1.21–1.52; $p = 1.40 \times 10^{-7}$) and FL risk (OR = 1.64; 95% CI = 1.45–1.86; $p = 5.0 \times 10^{-15}$) using independent cohorts (Karin E. Smedby et al., 2011). Additionally, a GWAS of 3857 DLBCL cases and 7666 controls used previously systemic lupus erythematosus (SLE) associated loci to evaluate the risk of DLBLC, finding HLA risk allele rs1270942. Another study evaluated multiple sclerosis (MS) and rheumatoid arthritis (RA) with DLBCL risk, but not genome-wide significance was reached (Bernatsky et al., 2017; Din et al., 2019). Moreover, HLA homozygosity was found to be associated with increased DLBCL risk for HLA-B, HLA-C and HLA-DRB1 alleles among Europeans (S. S. Wang et al., 2018).

GWAS have been also widely used for non-HLA alleles associations, more remarkably susceptibility risk have been found in different studies for: *PVT1* (rs4733601 and rs13255292) in three different GWAS studies in which one also found to be associated with MS risk ($p = 5 \times 10^{-8}$); *EXOC2* (rs116446171) (James R. Cerhan et al., 2014; Din et al., 2019; Park, Hong, Lee, & Koh, 2019) and *CD86* (rs2681416 and rs9831894) (James R. Cerhan et al., 2014; Kleinstern, Yan, et al., 2020). PVT1 is a non-coding RNA affecting MYC activation, a driver gene in lymphomas; EXOC2 functions at the interface between host defense and cell death regulation and CD86 is well known for its role in T-cell activation (Aslan et al., 2020; James R. Cerhan et al., 2014). Other, implicated loci are 2p23.3 (*NCOA1*), 3p24.1 (*EOMES-AZI2*), 5q31.3 (*ARAP3*) and 3q27 (*BCL6-LPP*); interestingly the BCL6 has been vastly documented to be involved in B-cell lymphomagenesis due to its role as critical regulator of germinal centers and rs6773363 (*EOMES-AZI2*) is indirectly involved in the activation of the NF-$\kappa\beta$ signaling pathway (Basso & Dalla-Favera, 2010; James R. Cerhan et al., 2014; Kleinstern, Yan, et al., 2020). Another study (399 DLBCL cases and 4243 controls) of Japanese population, identified risk for a variant within intron 3 of *CDC42BPB* (OR = 3.5; 95% CI = 2.13–5.88; $p = 3.30 \times 10^{-7}$), a gene with cell migration and cytoskeletal reorganization functions, a variant on *LNX2* (OR = 1.43; 95% CI = 1.23–1.67; $p = 6.57 \times 10^{-6}$), which indirectly mediates the NOTCH signaling and variant on *POU6F2* (OR = 1.57; 95% CI = 1.32–1.88; $p = 7.05 \times 10^{-7}$), a transcriptional regulator (Kumar et al., 2011). In addition to *PVT1*, other overlapping risk variants for DLBCL and MS were rs1270942 (*RDBP*), rs3130557 (*PSORS1C1*), and rs2425752 (*NCOA5*) (Din et al., 2019).

Another GWAS approach, using 491 DLBCL WGS data (31% discovery cohort; 69% validation cohort) and 1000 control WGS data, found NF-$\kappa\beta$ pathway activation by 3' cis-regulatory mutations on NFKBIZ but only on ABC DLBCL subtype which was later correlated with increasing expression on different DLBCL cell lines when compared to the non-mutated ones. GCB subtype, on the other hand, was associated with poor overall survival for *FCGR2B* over expressing patients (HR = 2.18; $p = 5.7 \times 10^{-3}$) (Arthur et al., 2018). Furthermore, though it has not been fully explored, some studies have shown that the presence of activation-induced cy-

tidine deaminase (AICDA) targeting motifs (WRC/GYW) within different point mutations, for example provoking induced translocations of PD-L1/PD-L2 with PIM1, TP63 and IGH or changes on the general mutational signatures across germinal center subtypes (Arthur et al., 2018; Georgiou et al., 2016; Muramatsu et al., 2000).

GWAS studies with survival data provided some evidence to finding potential prognostic or therapeutic targets in DLBCL, for example, a two-stage French study comprising four different cohorts in European population led to the discovery of two non-coding variants. The first one was rs7712513 at 5q23.2 (near *SNX2* and *SNCAIP*). The second one was rs7765004 at 6q21 (near *MARCKS* and *HDAC2*) that reached genome-wide significance for overall-survival association but not for progression free survival (PFS; Table 5.2). *SNX2* expression is reduced in human colorectal carcinoma and has been identified as a fusion partner of ABL1 in B-cell acute lymphoblastic leukemia; meanwhile, SNCAIP has been only reported in medulloblastoma studies. On the other hand, MARCKS has been widely studied for its role in invasion, proliferation, and drug resistance within different types of cancers (Duclos et al., 2017; Fong, Yang, & Chen, 2017; Ghesquieres et al., 2015; Y. Li et al., 2018). Another study using data recovered from the Genome Expression Omnibus (GEO) from 1804 DLBCL patients and performed a guilt-by-association analysis of only the 500 top-ranked CD20-associated gene probes. This study found *WEE1*, a replication checkpoint kinase that arrests cells at the G2/M checkpoint to give time for DNA repair, and *PARP1*, a repairing protein involved in high genomic instability and NF-$\kappa\beta$ activation, as potential candidates for DLBCL treatments. They further evaluated these targets using inhibiting drugs (AZD1775 for WEE1 and olaparib for PARP1) on different cell lines finding increased cytotoxic effects. Furthermore, a later study from the same group led to the discovery that combined WEE1 and anti-apoptotic protein inhibition enhances premature mitotic entry and DNA damage which may benefit genomic unstable DLBCL cells (Carrassa, Colombo, Damia, & Bertoni, 2020; Mathilde R. W. de Jong et al., 2018). Moreover, there are over 20 clinical trials exploring adavosertib, the most potent and selective WEE1 inhibitor, as a single agent or in combination for different indications (clinicaltrials.gov, October 2020).

FL

Follicular lymphoma is an indolent B-cell malignancy with higher five-year survival than DLBCL, though a subset of tumors can transform into more aggressive forms of lymphomas. FL is characterized by variable clinical outcomes, multiple relapses, and risk associations that includes family history of NHL (OR = 1.99; 95% CI = 1.55 to 2.54) and greater body mass index (OR = 1.15; 95% CI = 1.04 to 1.27 per 5 kg/m^2 increase) (Baecklund et al., 2014; Linet et al., 2014; C. F. Skibola et al., 2012). Two different three-stage GWAS studies in European populations, found that variant rs10484561 is associated with FL risk (OR_1 = 1.95; OR_2 = 1.64; 95% CI1 = 1.72–2.22; 95% CI2 = 1.45–1.86; $p < 1 \times 10^{-8}$) which, in addition, was later found to be implicated in DLBCL risk and in complete linkage disequi-

librium with the *HLADRB1*01:01 DQA1*01:01 DQB1*05:01* haplotype (LD-r^2 = 1.0). Additional *HLA-DQB1* variants associated with FL risk were, rs7755224 and rs2647012. The first one was in complete LD with variant rs10484561 suggesting that the effect was related to this. The last one was found 962 base pairs away from rs10484561. However, they were not in LD ($r^2 < 0.1$) and the association was protective and genome-wide significant after mutual adjustment (rs2647012-OR = 0.70; 95% CI = 0.67–0.78; $p = 4 \times 10^{-12}$; rs10484561-OR = 1.64; 95% CI = 1.45–1.86; $p = 5 \times 10^{-15}$), suggesting a totally different evolutionary origin (Conde et al., 2010; Karin E. Smedby et al., 2011). This last variant was later found on additional studies in different populations (Caucasians and Chinese) at genome-wide significance (James R. Cerhan et al., 2012; Qiao et al., 2013). Additional protective variants, found in another study with 699 cases and 2222 controls, were rs9275517 (OR = 0.63; 95% CI = 0.55–0.73; $p = 4.03 \times 10^{-11}$) and rs3117222 (OR = 0.66; 95% CI = 0.57–0.77; $p = 1.45 \times 10^{-7}$); furthermore, the second variant was correlated with higher *HLA-DPB1* expression in lymphoblastoid cell lines by using mRNA expression from MuTHER and Gen Cord datasets (Christine F. Skibola et al., 2012). In 2014, Skibola et al. identified two variants within the HLA-II class to be significantly associated with increased FL risk (rs12195582-OR = 1.78; 95% CI = 1.68–1.88; $p = 5.36 \times 10^{-100}$; rs17203612-OR = 1.43; 95% CI = 1.32–1.57; $p = 4.59 \times 10^{-16}$) (Christine F. Skibola et al., 2014). In addition to pleiotropy with DLBCL, HLA variants associated at genome-wide significance with FL has also been found for SLE, specifically two variants at *HLA-DOB* allele (rs1894406 and rs2071475) and one at *HLA-DRB1* allele (rs9271775) (Din et al., 2019). In respect to homozygosity, *HLA-DRB1* and *HLA-DRQ1* alleles were found to be associated with increased FL risk (S. S. Wang et al., 2018).

A two-stage study with 238 FL cases and 1233 controls from United States found a variant in *TAP2* gene (rs241447) to be associated with increase FL risk (OR = 1.82; 95% CI = 1.46–2.26; $p = 6.9 \times 10^{-8}$) but also with DLBCL (189 cases) risk, though DLBCL being not at genome-wide significance. TAP2 is part of the multidrug resistance protein (MRP)/TAP subfamily of ATP-binding cassette transporter, having an essential role for HLA class I protein loading on the cell surface and it is said that down-regulation or loss of function allows tumors to escape immune recognition (James R. Cerhan et al., 2012). One variant (rs6457327) near the psoriasis susceptibility locus (*PSORS1*) was found to be significantly associated to higher FL risk (OR = 1.69; 5% CI = 1.43–2.00; $p = 4.7 \times 10^{-11}$) among Europeans (Christine F. Skibola et al., 2009). Skibola et al. recompiled information from 22 studies (4523 cases and 13,344 controls) from European populations and found five significant associated loci: rs6444305 (OR = 1.21; 95% CI = 1.14–1.28; $p = 1.10 \times 10^{-10}$) located in *LPP* which encodes a LIM domain containing protein that has cell adhesion, migration and proliferation roles and also found 836.4 kb upstream of *BCL6*; rs13254990 (OR = 1.18; 95% CI = 1.11–1.24; $p = 1.06 \times 10^{-8}$) located intronic to *PVT1*, a frequent translocation site in aggressive B-cell lymphomas; rs4938573 (OR = 1.34; 95% CI = 1.26–1.46; $p = 5.79 \times 10^{-20}$) located 12.6 kb upstream *CXCR5*, involved in B-cell migration; rs4937362 (OR = 1.19; 95% CI = 1.13–1.25; $p = 6.76 \times 10^{-11}$) located near *ETS1*, a transcrip-

tion factor for B-cell differentiation; rs17749561 (OR = 1.34; 95% CI = 1.22–1.47; p = 8.28×10^{-10}) located near *BCL2*, an anti-apoptotic oncogene. Furthermore, another interesting, though not genome-wide significant, rs2681416 variant (near *CD86*) showed increased risk of FL (Christine F. Skibola et al., 2014). A Chinese study evaluating 792 cases and 1542 controls used additive genetic models adjusted with the false-positive rate probability to evaluate GWAS significance. This study found a variant on *IRF4* (rs872071), a crucial gene for B-cell development, to be associated with increased FL risk (Qiao et al., 2013). Moreover, this variant was found to be also associated with CLL risk in two additional European population studies (Crowther-Swanepoel, Broderick, et al., 2010; Di Bernardo et al., 2008). Additionally, five more variants (*CFB*, *MSH5*, *TNXB*, *LOC649925*, and *UBE2L3*) on chromosome 6 were associated with FL and SLE risk (Din et al., 2019).

ABC transporter variants associated with FL with worse PFS are *ABCA10* and *ABCA6* (rs10491178; HR = 3.17; 95% CI = 2.09–4.79; p = 5.24×10^{-8})which is in high LD ($r^2 > 0.8$) with another variant within the binding site of a transcription factor, *PAX5*, that has been correlated with aggressive subsets of B-cell NHL. These results were found among Europeans along with a variant on *CD46* (rs2466571; HR = 0.73; 95% CI = 0.58–0.91; p = 6×10^{-3}), *IL8* (rs4073; HR = 0.78; 95% CI = 0.62–0.97; p = 0.02), and *MTHFR* (rs1801131; HR = 0.59; 95% CI = 0.45–0.77; p = 1×10^{-4}), albeit positively associated with event-free survival after adjusting for age, sex and population stratification (Baecklund et al., 2014).

CLL

Another indolent lymphoma is CLL since the five-year survival rate is ~85% and it is characterized by a very rare incidence among Asian descendants compared to Caucasians and nearly double in males compared to females. Risk factors with CLL were previously identified to be family history of NHL (OR = 1.92; 95% CI = 1.42 to 2.61), hepatitis C virus infection (OR = 2.08; 95% CI = 1.23 to 3.49), and height (OR = 1.08, 95% CI = 1.00–1.17, p = 0.049) showing a slightly stronger trend among women (OR = 1.15, 95% CI: 1.01–1.31, p = 0.036). Conversely, immune function through allergy had a protective effect (OR = 0.87; 95% CI = 0.77 to 0.98) (Moore et al., 2020; S. L. Slager et al., 2014). Additionally, an analysis of 13 cancer types including 49,492 cancer case patients and 34,131 control patients found that individuals with a high risk score for CLL were at an increased relative risk of DLBCL (RR = 1.12, 95% CI = 1.07 to 1.16) (Sampson et al., 2015). HLA associations to CLL were mainly reported for the expanded haplotype *DRB4*01:01 DRB1*07:01 DQB1*03:03* in Caucasians (OR = 1.49; p = 1.79×10^{-7}), African Americans (OR = 28.03; p = 2×10^{-16}), and Hispanics (OR = 13.86; p = 9.59×10^{-9}) and *HLA-DRB4*0103* in a German study (RR = 2.74; p = 0.0025) (Gragert et al., 2014; Machulla et al., 2001). Other study from Caucasian population, found five variants from which two were associated with increase disease risk at genome-wide significance, one located near *HLA-DRB5* (rs674317) and the other near *HLA-DQA1* (rs9272535) (Susan L. Slager et al., 2011). On the other hand, so far there are no indicators of HLA zygosity

associations with CLL risk (Mueller & Machulla, 2002; S. S. Wang et al., 2018).

Because there have been several studies accessing CLL risk outside the HLA region, some variants have been validated across GWAS, for example variant rs17483466 (2q13; *ACOXL* and *BCL2L11*; $p_1 = 2.36 \times 10^{-10}$; $p_2 = 5 \times 10^{-9}$; $p_3 = 4 \times 10^{-17}$), rs735665 (11q24.1; *GRAMD1B*; $p_1 = 3.78 \times 10^{-12}$; $p_2 = 4 \times 10^{-24}$) and rs210142 (6p21.31; *BAK1*; $p_1 = 9.47 \times 10^{-16}$; $p_2 = 2.28 \times 10^{-16}$) (Sonja I. Berndt et al., 2013; Di Bernardo et al., 2008; Susan L. Slager et al., 2012; Speedy et al., 2014). Some biologically interesting variants include one encoding a protein involved in the signal transduction downstream of Ras, *PCEF1*, which happens to reside within a strong enhancer element (rs2236256; OR = 1.23; $p = 1.5 \times 10^{-10}$); a telomer protecting protein, *POT1* (rs17246404; OR = 1.22; $p = 3.40 \times 10^{-8}$) (Speedy et al., 2014); a member of the tumor necrosis factor which is essential for the signaling cascade in apoptosis, *ACTA2/FAS* (rs4406737; OR = 1.27; $p = 1.22 \times 10^{-14}$); a lymphocyte's apoptosis blocker, *BCL2* (rs4987855; OR = 1.47; $p = 2.66 \times 10^{-12}$) (Sonja I. Berndt et al., 2013); a member of the T-box gene family that regulates CD8+ T-cell differentiation and immunity, *EOMES* (rs9880772; $p_1 = 3 \times 10^{-11}$; $p_2 = \times 10^{-9}$; $p_3 = 2 \times 10^{-9}$), which is also critical during Fas deficiency for lymphoproliferation (Sonja I. Berndt et al., 2016; Law, Berndt, et al., 2017; Law, Sud, et al., 2017); a B-cell specific scaffold protein involved in B-cell antigen receptors, *BANK1* (rs71597109; OR = 1.17; $p = 1.37 \times 10^{-10}$); a master regulator of lymphocyte fate (B-cell vs T-cell) which is also involved in NOTCH pathway activation, *ZBTB7A* (rs7254272; OR = 1.17; $p = 4.67 \times 10^{-8}$) (Law, Berndt, et al., 2017) and a regulator of the PI3K/Akt pathway, *NCK1* (rs11715604; $p = 1.97 \times 10^{-8}$) (Law, Sud, et al., 2017).

Other implicated loci by GWAS include: 2q33.1 (rs3769825; *CASP10/CASP8*; $p = 2.5 \times 10^{-9}$), 2q37.1 (rs13397985; *SP140*; $p = 5.40 \times 10^{-10}$), 2q37.3 (rs757978; *FARP2*; OR = 1.39; $p = 2.11 \times 10^{-9}$), 3q25.2 (rs10936599; *MYNN*; $p = 1.74 \times 10^{-9}$), 4q25 (rs898518; *LEF1*; $p = 4.24 \times 10^{-10}$), 4q26 (rs6858698; *CAMK2D*; $p = 3.07 \times 10^{-9}$), 6p25.3 (rs872071 and rs9378805; *IRF4*; $p = 1.91 \times 10^{-20}$), 8q24.21 (rs2456449; $p = 7.84 \times 10^{-10}$), 11q24.1 (rs735665; *GRAMD1B*; $p = 3.78 \times 10^{-12}$), 12q24.13 (rs10735079; *OAS3*; $p = 2.34 \times 10^{-8}$), 15q23 (rs7176508; *DRAIC*; $p = 8 \times 10^{-18}$), 15q21.3 (rs7169431; *IRF8*; $p = 4.74 \times 10^{-7}$), 15q23 (rs7176508; $p = 4.54 \times 10^{-12}$), 16q24.1 (rs305061; *NEDD4* and *RFX7*; $p = 3.60 \times 10^{-7}$), and 19q13.32 (rs11083846; *PRKD2*; $p = 3.96 \times 10^{-9}$) (Sonja I. Berndt et al., 2016; Crowther-Swanepoel, Mansouri, et al., 2010; Di Bernardo et al., 2008; Sava et al., 2015; Susan L. Slager et al., 2012). Slager et al. found four additional IRF8 variants associated to both decreased CLL risk and increased IRF8 expression (using lymphocytes cell lines data) which is opposite to the previous finding by Crowther et al. one year earlier (Crowther-Swanepoel, Broderick, et al., 2010; Susan L. Slager et al., 2011). IRF4 and IRF8 are a strong finding due to its role as key regulator of B-cell development, proliferation, and lymphogenesis; furthermore, associations for variant (rs872071) were also found for FL (Crowther-Swanepoel, Broderick, et al., 2010; Qiao et al., 2013).

In spite of these findings several CLL associated risk variants have been also found for SLE, MS, and RA, only few have reached genome-wide significance; for

example, a variant on gene *BCL2* (rs4987855) which has anti-apoptotic activity (Din et al., 2019).

MZL

Marginal zone lymphoma, which comprises 10% of NHL cases, originate from marginal zone B cells present as three different types: extranodal MZL of mucosa-associated lymphoid tissue (EMZL) and splenic MZL (SMZL) and nodal MZL (NMZL) (Sonja I. Berndt et al., 2016; Susan L. Slager et al., 2012; Vijai et al., 2014, 2015). Risk factors for MZL include autoimmune conditions (EMZL OR = 6.40, 95% CI = 4.24–9.68; NMZL OR = 7.80, 95% CI = 3.32–18.33; SMZL OR = 4.25, 95% CI = 1.49–12.14), hepatitis C virus seropositivity (EMZL OR = 5.29, 95% CI = 2.48–11.28), self-reported peptic ulcers (EMZL OR = 1.83, 95% CI = 1.35–2.49), or family history NHL (NMZL OR = 2.82, 95% CI = 1.33–5.98). On the contrary, triglycerides levels were found as protective factor (OR = 0.90; 95% CI, 0.83–0.99) (Bracci et al., 2014; Kleinstern, Camp, et al., 2020; Speedy et al., 2014). GWAS studies in MZL, which are less extensive. One study comprised 1281 cases and 7127 controls of European ancestry in which a variant in HLA-B allele (rs2922994; OR = 1.64; 95% CI = 1.39–1.92; $p = 2.43 \times 10^{-9}$) was found to be associated with increased MZL risk taken together with a variant on *BTNL2* (rs9461741; OR = 2.24, 95% CI = 1.64–3.07; $p = 3.95 \times 10^{-15}$), a gene involved lymphocyte activation and antigen presentation (Susan L. Slager et al., 2012; Vijai et al., 2015). A second study assessing the role of HLA homozygosity in MZL risk (increased for HLA-B, HLA-C, and HLA-DRB1). A third study assessed the pleiotropy with SLE (*RDBP*, *PSORS1C1*, and *HLA-DQA1*) and RA (*CDH8*) (Din et al., 2019).

PCNSL

Recognized as mature post-germinal B cells (ABC subtype like), the central nervous system (CNS) DLBCL represents only $\leq$ 1% of all lymphomas and approximately 2% of all primary CNS tumors; furthermore, 95% of tumors have a comparative histology with systemic DLBCL (Sonja I. Berndt et al., 2016; Labreche et al., 2019). Since the blood brain barrier impedes R-CHOP treatment, high dose methotrexate (HD-MTX, >3 g/m^2) based regimens are the gold standard for PCNSL patients resulting in a five year survival rate of 30% and high risk of clinical neurotoxicity specially in patients > 60 (Garcilazo-Reyes et al., 2020; Law, Berndt, et al., 2017). Furthermore, PCNSL is less frequently associated with any atopic disorder (OR = 0.54, 95% CI = 0.33 to 0.87), but it is strongly associated with a family history of NHL (OR = 4.11, 95% CI = 1.58 to 10.66) and less clearly with lifetime cigarette exposure (OR = 1.51, 95% CI = 0.83 to 2.74, for 1–10 pack-years vs. nonsmoker) (J. R. Cerhan et al., 2014). The only study that has reported associations between genetic variants and PCNSL risk, evaluated 475 cases and 1134 controls from French population. This study found one variant at loci 6p25.3 (rs116446171; *EXOC2*; $p = 1.95 \times 10^{-13}$) previously found to be associ-

ated with DLBCL risk, other at loci 3p22.1 (rs41289586, *ANO10*, $p = 2.17 \times 10^{-8}$) and one strongly associated at HLA allele (rs2395192; between HLA-DRA and HLA-DRB5; $p = 1.81 \times 10^{-7}$). ANO10 is a calcium-activated chloride channel transmembrane protein that might be involved in the innate immune defense and indirect activation of the Ras/Raf/MEK/ERK signaling pathway which affects cell proliferation (Sonja I. Berndt et al., 2016; Labreche et al., 2019; Law, Sud, et al., 2017; C. Yu et al., 2020).

5.1.4 Conclusions

Initially data suggested increased risk of any DLBCL, FL, CLL, MZL, or PCNSL if family history of NHL was present, though specific genetic attributions for specific risk or prognosis had been lacking (Bracci et al., 2014; J. R. Cerhan et al., 2014; Linet et al., 2014; S. L. Slager et al., 2014). The decreasing cost and bioinformatics limitations for NGS and GWAS have augmented the ability to detect the genetic risk for NHL etiologies, pleiotropy with other diseases and, more importantly, clinical applications. Despite that many HLA alleles have been found to be not only associated to specific lymphoma subtypes but also within subtypes (DLBCL and FL) or with autoimmune diseases (MS, SLE, or RA), today they remain just informative for prognosis since no clinical use has been made (Din et al., 2019; Mills & Rahal, 2019; Zhong, Cozen, Bolanos, Song, & Wang, 2019). Peptide diversity reduction can increase tumoral escape from immune surveillance, which can be partially a consequence of HLA homozygosity, which was found to be a risk factor for most of the reviewed lymphomas (Mueller & Machulla, 2002; S. S. Wang et al., 2018).

On the other hand, GWAS findings outside HLA loci have led to the discovery of B-cell NHLs shared genetic risk with autoimmune diseases leading to finding of genes involved in cell cycle, apoptosis and telomere length, though studies are limited and total risk is still modest at genome wide-significance.

Taken together that hepatitis seropositivity is associated with DLBCL, CLL and MZL risk and that AICDA promiscuous off-target activity, highly present in lymphomas and induced by viral infection, can provoke important alterations (example translocations of *PD-L1/PD-L2* with *PIM1*, *TP63*, and *IGH* loci). Further efforts should be made to find correlations between these variables (Casellas et al., 2016; Georgiou et al., 2016; Kasar et al., 2015; Kohli et al., 2010). Future efforts should also be directed to extend studies among non-Caucasian populations, in order to clarify differences in susceptibility variants, and among B-cell NHLs subtypes since most studies have focused on DLBCL. Furthermore, there is an unmet need to translate theoretical information into clinical practice which has been done, for example, with the use of adavosertib, an WEE1 inhibitor, to increase response in DLBCL patients. In line with this, the incorporation of single-cell sequencing technology can help identify B-cell stages (dark/light zone) and cell cycle phases to further amplify the possibilities for therapy options (Holmes et al., 2020; Mathilde R. W. de Jong et al., 2018; Mathilde Rikje Willemijn de Jong et al., 2019).

5.1.5 Author Contributions

Conceptualization, I.H.-V., K.H.-X., and A.A.; methodology, I.H.-V., and M.B.; resources, I.H.-V., and K.L.; data curation, I.H.-V., K.H.-X., K.M., and A.A.; writing—original draft preparation, I.H.-V., K.L., and M.B.; writing—review and editing, all authors; visualization, all authors; supervision, I.H.-V., K.H.-X., and A.A; project administration, I.H.-V. and A.A.; funding acquisition, K.H.-X. and A.A. All authors have read and agreed to the published version of the manuscript.

5.1.6 Funding

This work was partially funded by the P DGOS/INCa PRT-K grant 2017–1-RT-04. A.A has also been supported by French Agence Nationale de la Recherche, as part of the second Investissements d'Avenir program (ANR-18-RHUS-0012) and by RAM Active Investments. I.H.-V. is supported by l'Association pour la Recherche sur les Tumeurs Cérébrales (ARTC).

5.1.7 Conflicts of Interest

The authors declare no conflict of interest.

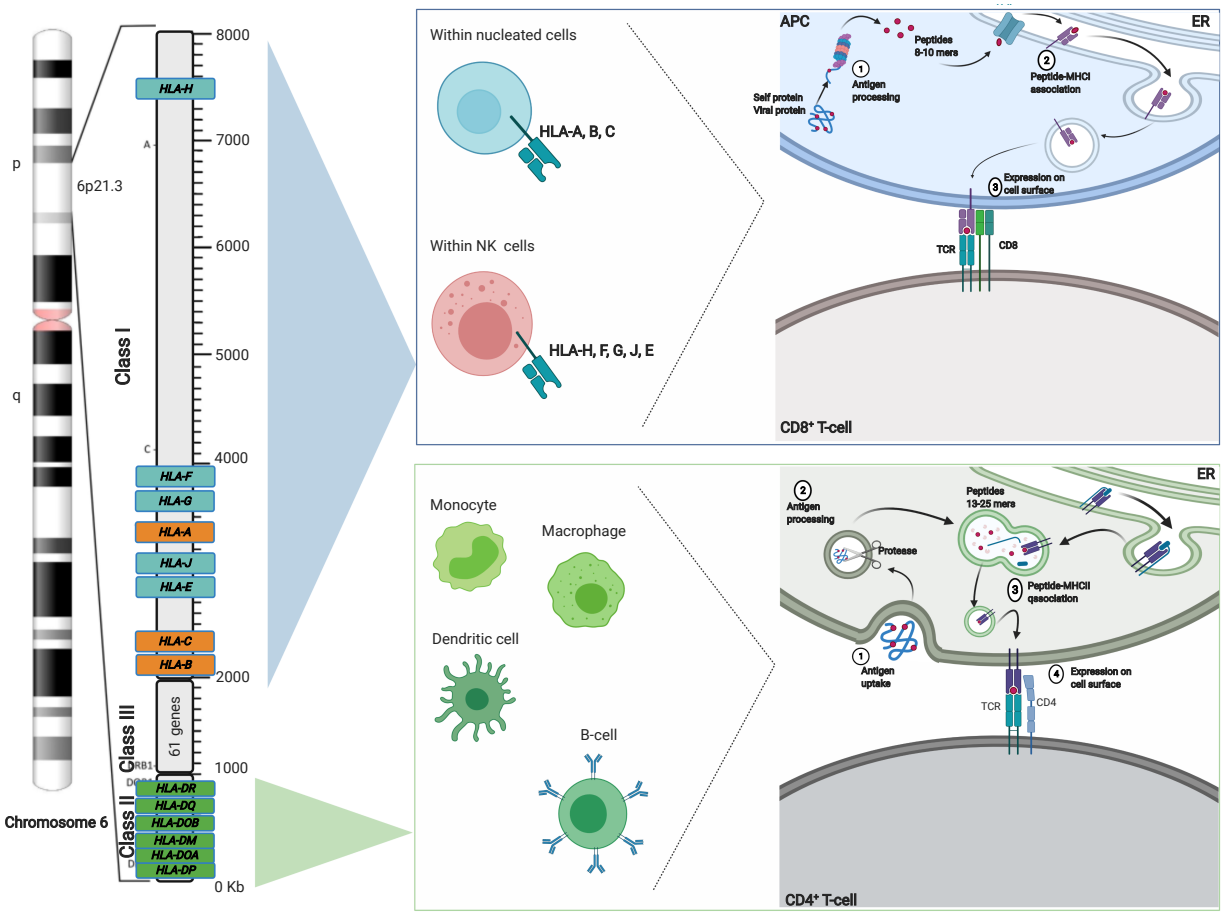


Figure 5.1: Schematic map of the human leukocyte antigen (HLA) genomic region showing the distribution of HLA genes along with the summarized mechanism of antigen presentation.

Table 5.1: Risk associations summary for diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), chronic lymphocytic leukemia (CLL), marginal zone lymphoma (MZL) and primary central nervous system lymphoma (PCNSL) with different loci identified by genome-wide association studies (GWAS).

Study	Year	Race / ethnicity	# cases/ #controls	SNP/alteration	Chr	Gene(s)	OR (95% CI)	p-value	Reference
DLBCL risk	2009	European	783/3,377	rs6457327	6p21.33	*PSORS1**	1.69 (1.43-2.00)	7.0×10^{-5}	[Skibola 2014]
	2011	European	1,592/6,581	rs10484561	6p21.32	*HLA-DQB1*	1.36 (1.21-1.52)	1.4×10^{-7}	[Smedby 2011]
	2011	Asian	399/4,243	rs751837	14q32	*CDC42BPB*	3.5 (2.127–5.88)	3.3×10^{-7}	[Kumar 2011]
				rs7097	13q12	*LNX2*	1.437 (1.23–1.67	6.5×10^{-6}	
				rs4551233	7	*POU6F2*	1.57 (1.32-1.88)	7.05×10^{-7}	
				rs4443228	4	--	2.43 (1.70-3.45)	7.03×10^{-7}	
	2013	Asian	1,328/6,930	rs6773854	3q27	*BCL6**, *LPP**	1.47 (1.32-1.65)	1.14×10^{-11}	[Law 2017, Tan 2013]
	2014	European	3,857/7,666	rs2523607	6p21.33	*HLA-B*	1.32 (1.21–1.44)	2.40×10^{-10}	[Cerhan 2014]
				rs116446171	6p25.3	*EXOC2**	2.20 (1.87–2.59)	2.33×10^{-21}	
				rs79480871	2p23.3	*NCOA1*	1.34 (1.21–1.49)	4.23×10^{-8}	
				rs13255292	8q24.21	*PVT1*	1.22 (1.15–1.29)	9.98×10^{-13}	
				rs4733601			1.18 (1.11–1.25)	3.63×10^{-11}	
				rs79464052	5q31.3	*ARAP3*	1.34 (1.21-1.49)	5.57×10^{-8}	
				rs2681416	3q13.33	*CD86*	1.16 (1.10-1.23)	8.17×10^{-8}	
	2015	Asian	1,124/3,596	rs116446171	6p25.3	*EXOC2**	2.04 (1.63–2.56)	3.9×10^{-10}	[Bassig 2015]
				rs13255292	8q24.21	*PVT1*	1.34 (1.19–1.52)	2.1×10^{-6}	
				rs2523607	6p21.33	*HLA-B*	3.05 (1.32–7.05	9×10^{-3}	
	2018	European	3,617/8,753	Homozygosity	6p21.33	*HLA-B, HLA-C*	1.31 (1.06-1.60)	8×10^{-4}	[Wang 2018]
				Homozygosity		*HLA-DRB1*	2.10 (1.24-3.55)	1×10^{-4}	
	2019	European	5,662/9,237	rs9831894	3q13.33	*CD86**, *ILDR1**	0.83	3.62×10^{-13}	[Kleinstern 2019]
			5,510/12,817	rs6773363	3p24.1	*EOMES**, *AZI2**	1.20	2.31×10^{-12}	
DLBCL-SLE	2017	European	3,857/7,666	rs4810485	20q13	*CD40*	1.09 (1.02-1.16)	0.013	[Bernatsky 2017]
				rs1270942	6p21.33	*HLA*	1.17 (1.01-1.36)	0.036	
	2019	European	3,617/46,436	rs1270942	6	*RDBP*	NA	5×10^{-8}	[Din 2019]
				rs3130557	6	*PSORS1C1*	NA		
				rs4733601	8	*PVT1*	NA		
DLBCL-MS				rs2425752	20	*NCOA5*	0.91	3.4×10^{-2}	
FL-risk	2009	European	645/3,377	rs6457327	6p21.33	*PSORS1**	1.69 (1.43-2.00)	4.7×10^{-11}	[Skibola 2014]
	2010	European	1,465/6,958	rs10484561	6p21.32	*HLA-DQB1*	1.95 (1.72-2.22)	1.12×10^{-29}	[Conde 2010]
				rs7755224	6p21.32		2.07 (1.76-2.42)	2.0×10^{-19}	
	2011	European	1,428/ 6,761	rs10484561	6p21.32	*HLA-DQB1*	1.64 (1.45-1.86)	5×10^{-15}	[Smedby 2011]
				rs2647012	6p21.32		0.70 (0.67-0.78)	4×10^{-12}	
	2012	Caucasians	699/2,222	rs9275517	6p21.32	*HLA-DRB1**	0.63 (0.55-0.73)	4.0×10^{-11}	[Skibola 2012]
				rs3117222		*HLA-DPB1**	0.66 (0.57-0.77)	1.45×10^{-7}	
	2013	Caucasians	238/1,233	rs2647012	6p21.32	*HLA-DQB1*	0.56 (0.45-0.69)	8.03×10^{-8}	[Cerhan 2012]
				rs241447	6p21.3	*TAP2*	1.82 (1.46-2.26)	6.9×10^{-8}	
		Asian	792/1,542	rs2647012	6p21.32	*HLA-DQB1*	1.20 (1.03–1.39)	0.018	[Qiao 2013]
				rs872071	6p25.3	*IRF4*	1.20 (1.05–1.38)	0.009	
	2014	European	4,523/13,344	rs12195582	6p21.32	HLA-DRB5	1.78 (1.68-1.88)	5.36×10^{-100}	[Skibola 2014]
				rs17203612	6p21.32	HLA-DRB1	1.43 (1.32-1.57)	4.59×10^{-16}	
				rs4938573	11q23.3	*CXCR5**	1.34 (1.26-1.43)	5.79×10^{-20}	
				rs4937362	11q24.3	*ETS1**	1.19 (1.13-1.25)	6.76×10^{-11}	
				rs6444305	3q28	*LPP*	1.21 (1.14-1.28)	1×10^{-10}	
				rs17749561	18q21.33	*BCL2**	1.34 (1.22-1.47)	8.28×10^{-10}	
				rs13254990	8q24.21	*PVT1**	1.18 (1.11-1.24)	1.06×10^{-8}	
				rs3751913	17q25.3	*CYBC1*	1.23 (1.14-1.33)	2.24×10^{-7}	
				rs2681416	3q13.33	*CD86*	1.16 (1.09-1.22)	2.33×10^{-7}	
				rs11082438	18q12.3	*SLC14A2*	1.33 (1.19-1.48)	4.01×10^{-7}	
	2018	European	2,686/8,753	Homozygosity	6p21.32	*HLA-DRB1*	1.54 (1.31-1.82)	1×10^{-4}	[Wang 2018]
				Homozygosity	6p21.32	*HLA-DQB1*	1.42 (1.23-1.65)	1×10^{-4}	
FL-DLBCL	2011	European	1,428/6,581	rs2647012	6p21.32	*HLA-DQB1*	1.36	1.4×10^{-7}	[Conde 2010]
FL-SLE	2019	European	2,686/46,436	rs1015166	6	*TAP2*	NA	5×10^{-8}	[Din 2019]
				rs1894406	6	*HLA-DOB*	NA		
				rs2071475	6	*HLA-DOB*	NA		
				rs2072634	6	*CFB*	NA		
				rs2293861	6	*MSH5*	NA		
				rs7774197	6	*TNXB*	NA		
				rs9271775	6	*HLA-DRB1*	NA		
				rs4938573	11	*LOC649925*	NA		
				rs7444	22	*UBE2L3*	NA		
CLL-risk	2001	European	101/157	---	6	*HLA-DRB4*0103*	2.74	2.5×10^{-3}	[Gragert 2014]
	2008	European	1,529/3,115	rs17483466	2q13	*ACOXL, BCL2L11*	1.39 (1.25–1.53)	2.36×10^{-10}	[Skibola 2009]
				rs13397985	2q37.1	*SP140**, *SP110**	1.41 (1.26–1.57)	5.40×10^{-10}	
				rs872071	6p25.3	*IRF4*	1.54 (1.41–1.69)	1.91×10^{-20}	
				rs9378805	6p25.3	*IRF4*	1.51 (1.38–1.65)	4.62×10^{-19}	

Table 5.1: Risk associations summary for diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), chronic lymphocytic leukemia (CLL), marginal zone lymphoma (MZL) and primary central nervous system lymphoma (PCNSL) with different loci identified by genome-wide association studies (GWAS).

			rs735665	11q24.1	*GRAMD1B*	1.45 (1.31–1.61)	$3.78x10^{-12}$	
			rs7176508	15q23	---	1.37 (1.26–1.50)	$4.54x10^{-12}$	
			rs11083846	19q13.32	*PRKD2*	1.35 (1.22–1.49)	$3.96x10^{-9}$	
2010	European	824/850	rs872071	6p25.3	*IRF4*	1.42 (1.23-1.63)	$9.96x10^{-7}$	[Di Bernardo 2008]
			rs735665	11q24.1	*GRAMD1B*	1.59 (1.34–1.88)	$1.23x10^{-7}$	
2010		2,503/5,789	rs757978	2q37.3	*FARP2*	1.39	2.11 x 10^{-9}	[Slager 2011]
			rs2456449	8q24.21	---	1.26	7.84 x 10^{-10}	
			rs7169431	15q21.3	*IRF8**	1.36	4.74 x 10^{-7}	
			rs305061	16q24.1	*NEDD4*, RFX7**	1.22	3.60 x 10^{-7}	
2011	Caucasians	690/1,295	rs305077	16q24.1	*IRF8*	0.66 (0.57-0.77)	3.37 x 10^{-8}	[Slager 2011]
			rs391525			0.64 (0.55-0.74)	3.16 x 10^{-9}	
			rs2292982			0.65 (0.56-0.75)	6.48 x 10^{-9}	
			rs2292980			0.66 (0.56-0.76)	1.89x 10^{-8}	
			rs615672	6p21.3	*HLA-DRB5*	1.42 (1.22-1.67)	1.29x 10^{-5}	
			rs674313			1.69 (1.41-2.01)	6.92x 10^{-9}	
			rs502771			1.61 (1.36-1.91)	5.58x 10^{-8}	
			rs9272219		*HLA-DQA1*	1.59 (1.34-1.90)	1.84x 10^{-7}	
			rs9272535			1.61 (1.35-1.92)	9.31x 10^{-8}	
2012	European/American	1,982/5,778	rs210142	6p21.33	*BAK1*	1.40 (1.25-1.57)	9.47x 10^{-16}	[Slager 2012]
2012	European/American	1,196/2,410	rs210142	6p21.33	*BAK1*	0.73 (0.68-0.79)	2.28x 10^{-16}	[Slager, Camp, 2012]
2013	European	3,100/7,667	rs4406737	10q23.31	*ACTA2*, FAS**	1.27 (1.19–1.33)	1.22x 10^{-14}	[Berndt 2013]
			rs4987855	18q21.33	*BCL2*	1.47 (1.32–1.61)	2.66x 10^{-12}	
			rs4987852			1.41 (1.27–1.56)	7.76x 10^{-11}	
			rs7944004	11p15.5	*C11orf21,* TSPAN32**	1.20 (1.13–1.27)	2.15x 10^{-10}	
			rs898518	4q25	*LEF1*	1.20 (1.14–1.27)	4.24x 10^{-10}	
			rs3769825	2q33.1	*CASP10, CASP8*	1.19 (1.12–1.25)	2.50x 10^{-9}	
			rs1679013	9p21.3	*CDKN2B-AS1*	1.19 (1.12–1.27)	1.27x 10^{-8}	
			rs4368253	18q21.32	*PMAIP1*	1.19 (1.12–1.27)	2.51x 10^{-8}	
			rs8024033	15q15.1	*BMF*	1.22 (1.15–1.30)	2.71x 10^{-10}	
			rs3770745	2p22.2	*QPCT*, PRKD3**	1.24 (1.15–1.33)	1.68x 10^{-8}	
			rs13401811	2q13	*ACOXL*, BCL2L11**	1.41 (1.30–1.52)	2.08x 10^{-18}	
2014	Europeans	3,748/8,574	rs10735079	12q24.13	*OAS3*	1.18 (1.12-1.26)	2.34x 10^{-8}	[Sava 2014]
2014	Europeans	2,883/8,350	rs2236256	6q25.2	*IPCEF1*	1.23 (1.15-1.30)	1.5x 10^{-10}	[Speedy 2014]
			rs10936599	3q26.2	*MYNN*	1.26 (1.17-1.35)	1.74x 10^{-9}	
			rs6858698	4q26	*CAMK2D*	1.31 (1.20-1.44)	3.07x 10^{-9}	
			rs17246404	7q31.33	*POT1*	1.22 (1.14-1.31)	3.40x 10^{-8}	
			rs1439287	2q13	*ACOXL*	1.37	5x 10^{-15}	
			rs13397985	2q37.1	*SP140*	1.43	5x 10^{-13}	
			rs872071	6p25.3	*IRF4*	1.39	3x 10^{-16}	
			rs735665	11q24.1	*GRAMD1B*	1.64	4x 10^{-24}	
			rs7176508	15q23	*DRAIC*	1.42	8x 10^{-18}	
			rs1044873	16.q24.1	*IRF8*	1.29	1x 10^{-9}	
2014	Caucasian	3,616/50,000	---	6	*HLA-DRB4*01:01*	1.49	1.79x 10^{-7}	[Gragert 2014]
	African-American	413/50,000		6	*DRB1*07:01*	28.03	$2x10^{-16}$	
	Hispanic	97/50,000		6	*DQB1*03:03*	13.86	$9.59x10^{-9}$	
2016	Europeans	5,058/13,197	rs9880772	3p24.1	*EOMES*	1.19 (1.13–1.25)	2.5x 10^{-11}	[Berndt 2016]
			rs73718779	6p25.2	*SERPINB6*	1.26 (1.16–1.36)	1.97x 10^{-8}	
			rs9815073	3q28	*LPP*	1.18 (1.11–1.25)	3.26x 10^{-8}	
			rs9308731	2q13	*BCL2L11*	1.19 (1.13–1.26)	1x 10^{-11}	
			rs10028805	4q24	*BANK1*	1.16 (1.10–1.22)	7.19x 10^{-8}	
			rs1274963	3p22.2	*CSRNP1*	1.18 (1.11–1.25)	2.12x 10^{-7}	
2017	Europeans	6,200/17,598	rs34676223	1p36.11	*MDS2*	1.19 (1.14-1.25)	5.04x 10^{-13}	[Law 2017]
			rs41271473	1q42.13	*RHOU*	1.19 (1.13-1.26)	1.06x 10^{-10}	
			rs71597109	4q24	*BANK1*	1.17 (1.11-1.22)	1.37x 10^{-10}	
			rs57214277	4q35.1	*MYL12BP2*, LINC02363**	1.13 (1.08-1.18)	3.69x 10^{-8}	
			rs3800461	6p21.31	*ILRUN*	1.20 (1.13-1.28)	1.97x 10^{-8}	
			rs61904987	11q23.2	*TMPRSS5*, DRD2**	1.24 (1.16-1.32)	2.46x 10^{-11}	
			rs1036935	18q21.1	*AC105227.1*, AC105227.2**	1.15 (1.10-1.21)	3.27x 10^{-8}	

Table 5.1: Risk associations summary for diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL), chronic lymphocytic leukemia (CLL), marginal zone lymphoma (MZL) and primary central nervous system lymphoma (PCNSL) with different loci identified by genome-wide association studies (GWAS).

				rs7254272	19p13.3	*ZBTB7A*, MAP2K2**	1.17 (1.10-1.23)	4.67x 10^{-8}	
				rs140522	22q13.33	*ODF3B*	1.15 (1.10-1.20)	2.7x 10^{-9}	
	2017	Europeans	1,842/7,324	rs11715604	3q22	*NCK1*	NA	1.97x 10^{-8}	[Law 2017]
				rs131821	22q13.33	*NCAPH2*	NA	7.49x 10^{-8}	
CLL-SLE	2019	European	2,492/46,436	rs10028805	4	*BANK1*	NA	$5x10^{-8}$	[Din 2019]
				rs1270942	6	*RDBP*	NA		
				rs17587	6	*PSMB9*	NA		
				rs3130557	6	*PSORS1C1*	NA		
				rs4987855	18	*BCL2*	NA		
				rs1439112	2	*MGAT5*	0.88	$4.7x10^{-2}$	
				rs10936599	3	*MYNN, ACTRT3, TERC, LRRC34*	0.86	$2.7x10^{-2}$	
				rs1317082	3			$1.5x10^{-2}$	
				rs13069553	3			$1.07x10^{-2}$	
				rs7621631	3			$1.8x10^{-2}$	
				rs10069690	5	*TERT*	1.16	$3.06x10^{-2}$	
CLL-MS				rs140522	22	*ODF3B*	0.90	$4.32x10^{-4}$	
				rs6793295	3	*LRRC34*	0.90	$1.24x10^{-2}$	
CLL-RA				rs3731714	2	*CASP10, PPIL3, CFLAR*	0.87	$4.69x10^{-2}$	
MZL-risk	2015	European	1,281/7,127	rs2922994	6p21.32	*HLA-B*	1.64 (1.39-1.92)	$2.43x10^{-9}$	[Mueller 2002]
				rs9461741	6p21.32	*BTNLA*	2.24 (1.64-3.07)	3.95×10^{-15}	
	2018	European	741/8,753	Homozygosity	6p21.33	*HLA-B*	1.34 (1.01-1.78)	0.012	[Wang 2018]
				Homozygosity	6p21.33	*HLA-C*	1.33 (1.04-1.70)		
				Homozygosity	6p21.33	*HLA-DRB1*	1.45 (1.05-1.91)		
MZL-SLE	2019	European	741/46,436	rs1270942	6	*RDBP*	NA	$5x10^{-8}$	[Din 2019]
				rs3130557	6	*PSORS1C1*	NA		
				rs532098	6	*HLA-DQA1*	NA		
MZL-RA				rs16947122	12	*FBXW8, HRK, TESC*	1.86	$3.35x10^{-2}$	
				rs1364229	16	*CDH8*	1.35	$1.10x10^{-3}$	
				rs7192064	16	*CDH8*	0.76	$4.36x10^{-2}$	
				rs2131402	16	*CDH8*	0.75	$1.01x10^{-2}$	
PCNSL-risk	2013	European	475/1,134	rs41289586	3p22.1	*ANO10*	3.82 (2.39-6.09)	2.17×10^{-8}	[Speedy 2014]
				rs116446171	6p25.3	*EXOC2**	4.99 (3.26-7.65)	$1.95x10^{-13}$	
				rs2395192	6p21	*HLA-DRA*, HLA-DRB5**	1.51 (1.29-1.76)	1.81×10^{-7}	

* Closest related gene; SLE: *lupus* erythematosus; MS: multiple sclerosis; RA: rheumatoid arthritis; NA: not available.

Table 5.2: Associations of different loci by GWAS with survival for DLBCL and FL

Study	Year	Race / ethnicity	# cases	SNP/alteration	Chr	Gene(s)	HR (95% CI)	p-value	Outcome	Reference
DLBCL	2015	European	1,537	rs7712513	5q23.2	*SNX2*, SNCAIP**	1.49 (1.29-1.72)	3.53×10^{-8}	↓ OS	[Ghesquieres 2015]
							1.39 (1.23-1.57)	2.08×10^{-7}	↓ PFS	
				rs7765004	6q21	*MARCK*, HDACS2**	1.47 (1.27-1.71)	5.36×10^{-7}	↓ OS	
							1.38 (1.22-1.57)	7.09×10^{-7}	↓ PFS	
	2018	European	210	---	1q23.3	*FCGR2B*	2.18	5.7×10^{-3}	↓ OS	[Arthur 2018]
FL	2014	European	586	rs10491178	17q24	*ABCA10*, ABCA6**	3.17 (2.09-4.79)	5.24×10^{-8}	↓ PFS	[Baecklund 2014]
				rs2466571	1q32.2	*CD46*	0.73 (0.58-0.91)	6×10^{-3}	↑ EFS	
				rs4073	4q13.3	*IL8*	0.78 (0.62-0.97)	0.02	↑ EFS	
				rs1801131	1p36.22	*MTHFR*	0.59 (0.45-0.77)	1×10^{-4}	↑ EFS	

* Closest related gene; OS: overall survival; PFS: progression free survival; ↓: inferior; EFS: event free survival; ↑: superior

5.2 Pan-cancer landscape of AID-related mutations, composite mutations and their potential role in the ICI response

Isaias Hernández-Verdin[1], Kadir C. Akdemir[2], Daniele Ramazzotti[3], Giulio Caravagna[4], Karim Labreche[1], Karima Mokhtari[1,5,6], Khê Hoang-Xuan[1,7], Matthieu Peyre[1,8], Franck Bielle[1,5,6], Mehdi Touat[1,7], Ahmed Idbaih[1,7], Alex Duval[9], Marc Sanson[1,6,7], Agusti Alentorn[1,7,*]

[1]Sorbonne Université, Inserm, CNRS, UMR S 1127, Institut du Cerveau et de la Moelle épinière, ICM, Paris, France

[2]Departments of Genomic Medicine and Neurosurgery, University of Texas MD Anderson Cancer Center, Houston, TX, USA

[3]Department of Medicine and Surgery, University of Milano-Bicocca, Milano, Italy

[4]Cancer Data Science Laboratory, Dipartimento di Matematica e Geoscienze, Università degli Studi di Trieste, Italy

[5]Department of Neuropathology, Pitié Salpêtrière-Charles Foix, Paris, France.

[6]Onconeurotek, AP-HP, Hôpital Pitié-Salpêtrière, Paris, F-75013, France.

[7]Department of Neurology-2, Pitié-Salpêtrière University Hospital, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France.

[8]Department of Neurosurgery, AP-HP, Hôpital Pitié-Salpêtrière, Paris, F-75013, France.

[9]Sorbonne Université, INSERM, Unité Mixte de Recherche Scientifique 938 and SIRIC CURAMUS, Centre de Recherche Saint-Antoine, Equipe Instabilité des Microsatellites et Cancer, Equipe labellisée par la Ligue Nationale contre le Cancer, F-75012 Paris, France; Sorbonne Université, Genetics Department, AP-HP.Sorbonne Université, hospital Pitié-Salpêtrière, F-75012 Paris, France

[*]Corresponding author:

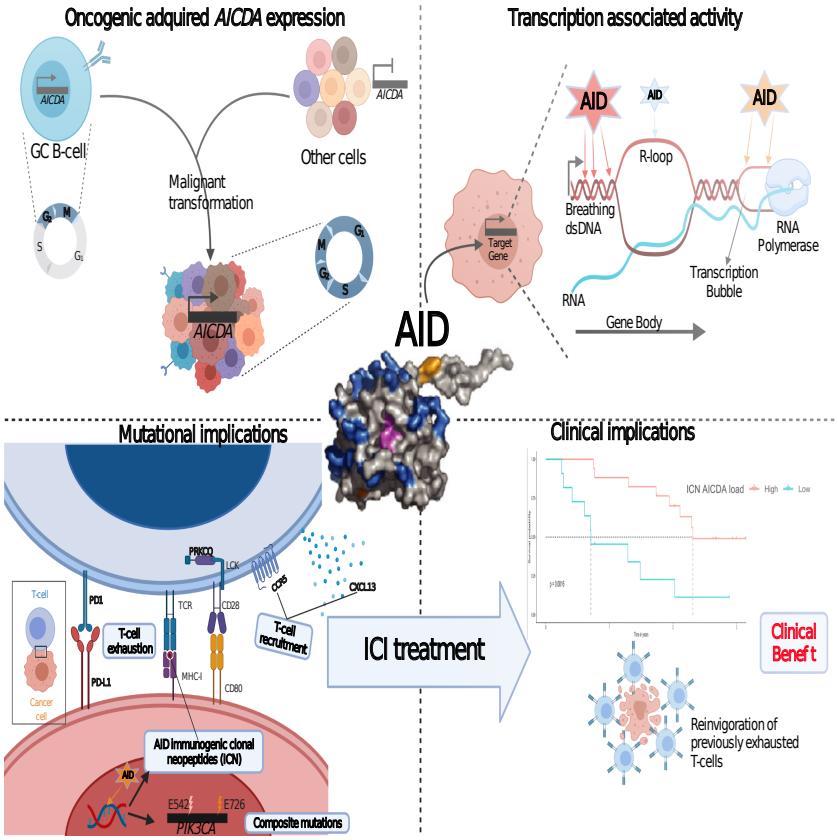
Agustí Alentorn, MD PhD

Email address: agusti.alentorn@icm-institute.org

Mailing address: Department of Neurology-2, Mazarin, Groupe Hospitalier Pitié Salpêtrière, 75013 Paris, France

Phone: +33 142164160

Graphical abstract



In Brief

A combined bulk and single cell multi-omic analysis of over 50,000 patients and 2.5 million cells across 80 tumor types reveals oncogenic acquired AICDA expression inducing composite mutations and clonal immunogenic neoepitopes that are associated with favorable outcome in patients treated by immune-checkpoint inhibitors.

Highlights

- Pan-cancer analysis of AID mutations using > 50,000 samples, 2,000 ICI treated cases and 2.5 million cells with genome, exome and transcriptome data
- Oncogenic transient AICDA expression induces mutations mainly during transcription of its off-target genes in virtually all cancers
- AID is implicated in composite mutations on weakly functional alleles and immunogenic clonal neoepitopes at hotspots with greater positive selection
- AID mutational load predicts response and is associated with favorable outcome in ICI treated patients

Abstract

Activation-induced cytidine deaminase, AICDA or AID, is a driver of somatic hypermutation and class-switch recombination in immunoglobulins. In addition, this deaminase belonging to the APOBEC family may have off-target effects genome-wide, but its effects at pan-cancer level are not well elucidated. Here, we used different pan-cancer datasets, totaling more than 50,000 samples analyzed by whole-genome, whole-exome or targeted sequencing. AID synergizes initial hotspot mutations by a second composite mutation. Analysis of 2.5 million cells, normal and oncogenic, revealed AICDA expression activation after oncogenic transformation and cell cycle regulation loss. AID mutational load was found to be independently associated with a favorable outcome in immune-checkpoint inhibitors (ICI) treated patients across cancers after analyzing 2,000 samples. Finally, we found that AID-related neoepitopes, resulting from mutations at more frequent hotspots if compared to other mutational signatures, enhance CXCL13/CCR5 expression, immunogenicity, and T-cell exhaustion, which may increase ICI sensitivity.

Keywords: AICDA; AID; pan-cancer; single-cell; composite mutations; clonal; immunotherapy; immune-checkpoint inhibitors; neoepitopes.

5.2.1 Introduction

Naive B-cells enter the germinal centers (GC) of secondary lymphoid organs after being activated by a cognate antigen, where they induce the production of Activation-induced cytidine deaminase (AICDA), especially during the G2-M phases of the cell cycle (Figure 5.2A, I).

AID (encoded by *AICDA*) is involved in the diversification of the variable (V) or switch domains of immunoglobulin (IG) genes during the G1-S phases of the cell cycle. It is responsible for somatic hypermutation (SHM) in the dark zone of the GC and class switch recombination (CSR) in the light zone (Figure 5.2A, II) (Honjo, Kinoshita, & Muramatsu, 2002; Honjo, Muramatsu, & Fagarasan, 2004; Q. Wang et al., 2017). AID deamination of cytosine to uracil also occurs during IG gene transcription and inside particular DNA patterns (Figure 5.2A, III) (Branton et al., 2020). Mutations can arise as A->C at WA motifs ($W = A/T$) when resolved by the error-prone DNA polymerase-η, which has been defined as non-canonical AID (COSMIC signature 9), or as C->T/G at WRCY motifs ($R = purine; Y = pyrimidine$) when resolved by base excision repair (BER) or mismatch repair (MMR) pathways, which has been defined as canonical-AID (c-AID, Figure 5.2A, III) (Delgado et al., 2020). Although the single-base substitution (SBS) COSMIC somatic signatures SBS84 and SBS85 (v3.2) have recently been linked to c-AID activity, they were discovered in a trinucleotide context (specifically at RCY motifs), which does not always correspond to the observed tetranucleotide context in which c-AID acts (WRCY motifs) (Australian Pancreatic Cancer Genome Initiative et al., 2013; Kasar et al., 2015; PCAWG Mutational Signatures Working Group et al., 2020). Furthermore, AID belongs to the same enzyme family as APOBEC3A and APOBEC3B, which are known to be a source of somatic mutations in a variety of malignancies and are designated by the SBS2 and SBS13 signatures according to Alexandrov, but unlike c-AID, act in trinucleotide context (TCW motifs) (Australian Pancreatic Cancer Genome Initiative et al., 2013; Roberts et al., 2013; Swanton, McGranahan, Starrett, & Harris, 2015). Off-target AID activity has also been reported in lymphomas and other hematological cancers (Pasqualucci et al., 2008; Rustad et al., 2020), but only in a few solid tumors (Komori et al., 2008; Sapoznik et al., 2016; Sawai et al., 2015; Shimizu et al., 2014). Despite this, no detailed characterization of the involvement of AID-related mutations at the pan-cancer level, as well as their potential mutational and clinical implications, has been performed. To test this, we analyzed 18 tumor types spanning 41 research, as well as three studies covering the human and mouse of normal cells, all at single-cell resolution, to show that AICDA expression is acquired during malignant transformation (Han et al., 2018, 2020; Zheng et al., 2017). The c-AID mutations were then characterized across 49 thousand tumoral samples (9 human cohorts and 3 non-human cohorts, see STAR

methods), revealing that: i) they are found at a frequency of 5.2% (5.1-5.3%) in virtually all cancers (human and non-human); ii) they show stronger activity at transcriptionally active domains; and iii) they synergize initial hotspot mutations by a second composite mutation.

Additionally, since the APOBEC mutational signature (SBS2 and SBS13) has been proposed as a biomarker for ICI response in some cancers (Litchfield et al., 2021; S. Wang, Jia, He, & Liu, 2018), we used more than 2.000 ICI-treated samples (Miao et al., 2018; Pender et al., 2021; Samstein et al., 2019), finding AID-related fraction of mutations as an independent prognostic value to ICI after adjusting by TMB and APOBEC signature.

Overall, we used more than 50.000 samples covering more than 80 tumor types at the bulk level and close to 2.5 million cells at single-cell resolution to thoroughly describe the landscape of AID-related mutations (see Figure 5.2, Figure 5.3A, and Supplementary Tables 1-2).

5.2.2 Results

AICDA expression is activated under oncogenic conditions, according to scRNA studies

The first step we took was to improve the characterization of *AICDA* expression across normal tissues and cells by analyzing scRNA-seq data from ~ 600, 68, and 350 thousand cells from the human cell landscape (HCL), peripheral blood mononuclear cells (PBMCs), and the mouse cell atlas (MCA), respectively (Han et al., 2018, 2020; Zheng et al., 2017). *AICDA* expression was observed primarily in adult epityphlon and adult pleura (Supplementary Figure 1A) in human samples and adult small intestine, ovary, pleura, and spleen in mouse samples when examined by tissue (Supplementary Figure 1B).

The expression was significantly stronger in adult stages than in fetal or embryonic stages (Supplementary Figure 1C). Using the cell type annotation, in the HCL, we observed the highest expression in B-cells followed by fibroblasts (Supplementary Figure 1A). Further analysis of only PBMCs led to the finding that *AICDA* is also expressed in CD8 T-cells and induced regulatory T-cells, but at lower levels than that of B-cells (Figure 5.2B) as expected. Interestingly, the expression of the *BER* gene *UNG* and the MMR genes *MSH2/MSH6*, involved in downstream reparation of AID mutations, were expressed altogether with *AICDA* only within B-cells, CD8 T-cells, and induced regulatory T-cells (Tregs).

The next steps were addressing whether or not *AICDA* is expressed under oncogenic conditions, deciphering the cell subtypes' contribution to *AICDA* expression, and evaluating its regulation through the cell cycle. We gathered information from 41 oncogenic single-cell studies comprising around 1.5 million cells and 18 tumor

types (see STAR Methods) and found that malignant cells express *AICDA* across all but one tumor type basal cell carcinoma (BCC), being stronger within skin cutaneous melanoma (SKCM), medulloblastoma (MB), non-small cell lung cancer (NSCLC), non-small cell lung cancer (HNSC), and glioma. Furthermore, the expression was lost when tumoral cells were removed from the samples, and it was not found in Tregs. Interestingly, *AICDA* is expressed by a fraction of immune population cells, including B-cells and fatigued CD8 T-cells in SKCM, diffuse large B cell lymphoma (DLBCL), and NSCLC, as well as monocytes/macrophages in some SKCM and glioma studies, as well as fibroblasts and endothelial cells from the stromal population (Figure 5.2C). Surprisingly, its expression was observed across all cell cycle phases. However, it was slightly higher at G2/M; on the other hand, BER and MMR related genes are markedly more expressed at the S phase, meanwhile, the expression of nonhomologous end joining (NHEJ) related genes (*NHEJ1*, *TP53BP1*, or *Trp53bp1* in mouse) remained practically unchanged. Additionally, the expression was higher in SKCM, DLBCL and gliomas independently of treatment (Figure 5.2D). Altogether, these results suggest that cells activate *AICDA* expression after malignant transformation without cell cycle regulation and whose levels vary upon tumor type.

Landscape of AID-related mutations at pan-cancer level

Given that *AICDA* is expressed in different cancer types, we next identified the mutations induced by c-AID activity by tracking the C to G/T mutations within its specific WRCY motifs. We discovered AID-related mutations in the great majority of malignancies investigated while evaluating the PCAWG data (ICGC; 2,775 cancer patients and 35 cancer types) (Figure 5.3). Overall, AID-related mutations were detected in 5.2% (5.1−5.3% at 95% confidence interval [CI]), while APOBEC mutations (SBS2+SBS13) were found in 6.5% (6.4−6.7% at 95% CI; Figure 5.3B). Using the TCGA, MSKCC cohorts, and various pediatric datasets, we observed similar results at the pan-cancer level (Supplementary Figures 2-3). Conversely, as expected, the frequency of AID-related mutations was slightly higher in hematological cancers at approximately 8% (Supplementary Figure 3D). Intriguingly, the AID mutations were also identified in canine melanoma, glioma, and osteosarcoma at a frequency of 6.0%, 4.7%, and 2.9%, respectively (Supplementary Figure 3E). Moreover, to discard an association of our tetranucleotide-based c-AID mutations (using ICGC cohort) with other COSMIC somatic signatures (v3.2) we computed the cosine similarity scores and observed SBS84, SBS9, and SBS85 showing low cosine scores of 0.497, 0.157, and 0.039, respectively. Next, we simulated each sample's mutations 1,000 times, while maintaining the mutational patterns at pentanucleotide resolution (SBS-1536) and mutation load, to generate a distribution of mutations and a null hypothesis about the number of c-AID related mutations

generated by chance (globally and per tumor type). We observed a 2.69 (2.67−2.71 at 95% CI) enrichment of observed versus expected c-AID related mutations globally, and only 3/2727 samples having significantly more c-AID mutations by chance (two-sided Fisher exact test; Supplementary Figure 4). These observations indicate that it is very unlikely that the majority of the observed c-AID mutations are the result of chance or an already reported mutational signature.

Concerning the genomic distribution of AID motifs in the normal genome, the quantity is not different across chromosomes when adjusting the motifs' number by chromosome length (FDR corrected p-value Wilcoxon-test; Supplementary Figure 5A-B). Regarding AID-related mutations, in DLBCL most commonly affected chromosomes involved the presence of either immunoglobulin related genes: *IGH* (chr14), *IGL* (chr22), *IGK* (chr2), or genes already related with off-target AID activity: *PIM1*, *IRF4*, *HIST1H1C* (chr6; Supplementary Figure 5C-F) (Lossos, Levy, & Alizadeh, 2004). Globally speaking, for the majority of tumor types the highest density of AID mutations were located in chromosome 5, in which *GPR98* and *DNAH5* were frequently affected, followed by chromosomes 17 and 2 (Supplementary Figures 6-9).

Interestingly, within the driver genes context, hematological cancers (i.e. non-Hodgkin's lymphoma (Lymph-BNHL), DLBCL) and MB had the highest signature contribution of AID provoked mutations. Furthermore, among the involved targets, *TP53*, in all cohorts; *IDH1*, in hematological cancers, GBM and LGG; and *PIK3* genes (TCGA and ICGC cohorts), were recurrently altered (Supplementary Figure 10). These results were also confirmed using a selection intensity approach of every somatic mutation within the ICGC dataset, showing a higher selection intensity of *PIK3CA*, *NFE2L2* but also in "minor" *IDH1* mutations (i.e. not R132H) and *PTEN* (Figure 5.3C).

AICDA expression and AID-related mutations were not correlated, and only in thyroid cancer (THCA) were slightly positively correlated ($Rho = 0.18$, $p_{adj} = 0.01$), suggesting that *AICDA* is not constitutively activated in any cancer. The AID mutations were more frequently negatively correlated with the tumor mutation burden (TMB) of cancers from TCGA (i.e. in adenoid cystic carcinoma (ACC), kidney renal papillary cell carcinoma (KIRP), kidney renal clear cell carcinoma (KIRC), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), ovarian cancer (OV), and THCA; Supplementary Figure 11). In brief, we found AID activity leaves important DNA footprints across human and not-human tumors, including driver genes.

Relationship of AID-related mutations with immune-related signatures and the presence of viral genome

Taking together that normal *APOBEC* expression is induced after viral exposition, recent findings of oncogenic *APOBEC* expression correlating with positive human papillomavirus (HPV) infection in HNSC (Cannataro et al., 2019) and studies showing a correlation between AID-related mutations with chronic infections (i.e. not viral), like with Helicobacter pylori (H. pylori) in precancerous stages of stomach cancer or Plasmodium infection in B cell lymphoma (Robbiani et al., 2015; Shimizu et al., 2014), we sought to analyze the potential relationships between the presence of AID-related mutations, the *AICDA* expression and the presence of different oncogenic viruses at pan-cancer level using the TCGA dataset (Figure **??**A-B). In addition, we also analyzed their relationship with different immune-related cells, obtained by deconvolution. The AID-related mutations only showed significant enrichment in stomach adenocarcinoma (STAD) tumors with negative Epstein-Barr virus (EBV) infection (Figure **??**B). Intriguingly, *AICDA* expression was significantly higher in bladder cancer (BLCA), HNSC, and cervical squamous cell carcinoma (CESC) with HPV infection (2.68 versus 2.52; 2.62 versus 2.51; 2.50 versus 2.45; respectively, Wilcoxon-test, Figure **??**A). We did not find other remarkable differences in *AICDA* gene expression or when we considered the AID-related mutations (Figure **??**), suggesting that *AICDA* is rarely constitutively expressed and is rather transient at pan-cancer level. In some cancer types, we discovered a link between AID-related mutations and the presence of M1 macrophages, T CD4, and CD8 cell populations (Figure **??**C; Spearman correlations with FDR-adjusted p-values). Furthermore, the existence of these immune cell populations was co-expressed in the same cancer types in the same direction in the vast majority of cases (Figure **??**C). *AICDA* gene expression, on the other hand, was not co-expressed with these three immune cell types but was positively correlated with B cell naive and different properties of B cell receptor (BCR), as well as a lymphocyte cell infiltration signature (Supplementary Figure 12; Spearman correlations with FDR-adjusted p-values).

This suggests that only HPV infection, in general, does trigger *AICDA* expression but does not correlate with AID-related mutations.

AID and APOBEC activity is higher at transcriptionally active domains but their relation with the MMR activity is contrariwise

Initial studies have revealed that the mutation frequency is increased in late-replicating regions (G2/M phases), mainly, due to increased MMR activity on early zones (G1/S phases) (Tomkova, Tomek, Kriaucionis, & Schuster-Böckler, 2018). However, recent works have found 3D chromatin organization to be bet-

ter correlated with the mutational load than with the replication time alone but the direction, towards active or inactive domains, is shaped by the mutational signature (Akdemir et al., 2020). By using replication timing alone we found that the AID and SBS2 signatures have a clear late replication enrichment (global p-values = $3.09e^{-35}$ and $3.83e^{-3}$, respectively; Supplementary Figure 13). However, the enrichment zone changes across tumor types, being more "early" in Bladder transitional cell carcinoma (TCC), Cervix squamous cell carcinoma (SCC), Uterus adenocarcinoma (AdenoCa), thyroid adenocarcinoma (Thy-AdenoCA), Lung cancers, and others which are mostly already reported APOBEC-prone cancer types (Supplementary Figure 14A-B). On the other hand, the SBS13 signature is not enriched globally (p-value = 0.88) but it is in the early zone for the same tumors as SBS2, which is consistent with previous reports (Supplementary Figure 14C) (Morganella et al., 2016; Tomkova, Tomek, Kriaucionis, & Schuster-Böckler, 2018). In B-cells AID is mostly active in G0/G1 phase and has been proved to induce mutations before malignant transformation (Kasar et al., 2015). Alternatively, our findings indicate that it might change to be able to induce mutations at late replicating zones on hematological cancers and others (kidney renal cell carcinoma (Kidney-RCC), pancreatic neuroendocrine tumors (Panc-Endocrine), and Stomach adenocarcinoma; Supplementary Figure 14A), which is in line with our previous findings of the cell cycle regulation loss for *AICDA* expression. Next, we used topologically associated domains (TADs) boundary information of active and inactive domains, in terms of transcription, to see the distribution of AID/APOBEC mutations across chromatin folding domains (Akdemir et al., 2020). We found AID mutations occurring more towards active domains than inactive ($FC = 3.63$; p-val $= 5.01 \times 10^{-98}$), especially at the TADs boundaries (Figure 5.5A). As previously described, we found that APOBEC signatures are also causing mutations towards active domains but the active/inactive ratio is notably higher for the SBS13 than the SBS2, indicating distinct molecular underpinnings (Supplementary Figure 15).

Meanwhile, in normal B-cells, the MMR is in charge of repairing the AID-canonical related mutations, thus reducing its mutagenesis; within the oncogenic context, previous studies have demonstrated that it is the MMR machinery itself that can increase mutagenesis indirectly because when it is repairing any mutation it creates 800 bp fragments that are targets of APOBEC mutagenesis (Mas-Ponte & Supek, 2020). To further test these reported observations at pan-cancer scale, we attempted to see the differences of AID/APOBEC mutagenesis between MSI or MSS tumors; additionally, considering that MMR activity has been reported to be higher on early replicating zones (Tomkova, Tomek, Kriaucionis, & Schuster-Böckler, 2018), we separated mutations falling within early or late replicating zones. We observed that tumors having impaired MMR machinery (MSI samples)

had higher AID mutations in early replicating zones compared to MSS tumors (p-value = 0.0043; Wilcoxon test), but there was no significant difference in AID mutations falling in late replicating zones (MSI vs MSS; p-value = 0.3029). Likewise, when we compared early vs late within MSI tumors we found higher AID mutations falling in early (p-value = 0.0059) but the opposite for MSS tumors (p-value = $2e^{-16}$; Figure 4B). Furthermore, we also validated this hypothesis by analyzing 19,936 additional tumors (MSKCC cohort) finding a significant increment in the global number of AID mutations in MSI tumors or APOBEC mutations in MSS tumors (p-value = $2e^{-16}$ & $6e^{-10}$; Wilcoxon test; Supplementary Figure 16A).

Therefore, our analyses suggest that both APOBEC and AID mutations have a higher preference at transcriptionally active domains, though the effect is more marked for AID. Moreover, as expected, AID mutations are repaired by the MMR machinery while APOBEC mutations are enhanced by the MMR activity.

Oncogenic AID activity differs according to transcription direction

AID activity within the normal context takes place especially during transcription elongation, when the polymerase becomes stalled, and requires a licensing step to regulate over-activity which can be bypassed when abnormal high nuclear levels of AID are present. On the other hand, R-loops, a hybrid structure of B-form double-stranded DNA and A-form dsRNA, are formed during transcription which increases DNA exposure and has been linked to AID activity (Ginno, Lott, Christensen, Korf, & Chédin, 2012; Methot et al., 2018). By using genomic coordinates of R-loop associated regions and the ICGC cohort (WGS data), we attempted to answer if, within the tumoral context, AID mutations were more localized in or out these regions compared to either APOBEC mutations (COSMIC SBS2/SBS13) or other mutations. Surprisingly, only 0.18% (1130/629,871) of all AID mutations were "in R-loops" which was not significantly different to those caused by SBS2 (0.19%; 456/241,695; p-value = 0.37; two-sided Fisher exact test) or SBS13 (0.19%; 400/204,922; p-value = 0.15) (Supplementary Table 3). Overall, this suggests that within the oncogenic context, AID promiscuous activity is not related to R-loop formation.

Since the R-loop forming regions do not cover all the transcription start sites (TSS), we next analyzed the AID mutation's distribution around the TSS as previous studies showed recruitment of AID to those sites (Methot et al., 2018). By dividing mutated genes based on strandness, we found a very particular pattern for the AID mutations falling on the negative strand compared to the positive strand. Mutations accumulate near the TSS and towards the gene body while maintaining a more constant mutational load compared to the opposite direction, the positive strand (Figure 5.5C), or even if compared to APOBEC signatures at either strand (Supplementary Figure 16B and C).

Next, we wondered if the AID mutations of a specific gene were produced during transcription of that same gene. To answer this we used 1,130 samples (comprising 24 tumor types) from which the mutations and expression data were available (ICGC cohort) and correlated the number of AID induced mutations occurring in gene i (AID_Muts_{gi}) to the expression of the same gene i (Exp_{gi}) within each tumor type since the expression and mutations vary greatly in this context. Only 6.6% of the total different mutated genes (21,341) were expressed from which 6.0% (81/1,403, not repeated genes) were correlated with its corresponding gene expression per tumor type (p.adj < 0.05, Spearman $Rho > 0$); additionally, more than one-third of these genes were found in hematological cancers (Lymph-BNHL and Lymph-CLL). Gene set enrichment analysis revealed immunoglobulin V region-related genes (adjusted p-value $= 1.2 \times 10^{-11}$) within hematological cancers (Supplementary Figure 16D). Altogether, our analysis suggests that AID activity is coupled to the transcription process with immunoglobulin genes in hematological cancers following the line of "normal" context as the expression is more constitutive. However, the mutations in other genes are probably produced during short-term transcription of both the affected gene and *AICDA* whose dynamics depend on the strand location of the gene and hence the direction of transcription.

AID temporal mutations vary across tumor types and are independent of replication timing

Point mutations, within a cell, arising before a chromosomal locus duplication will give rise to different cell lineages compared to the mutations happening after the duplication. WGS data can be used to infer the number of allelic copies and hence the ratio of duplicated to non-duplicated mutations within a gained region can be used to estimate the time point when the gain happened and define variants as clonal (happening at early points) or as subclonal (happening at more recent time points) (PCAWG Mutational Signatures Working Group et al., 2020). By combining 13 million point mutations (SNVs only) and available copy number variations (CNV) data from the PCWAG dataset (2,707 samples), we evaluated the molecular timing of AID provoked mutations and compared them to APOBEC signatures (SBS2 and SBS13) and the hypermutation attributed signature (SBS9, non-canonical AID). We found that 90.5% of mutations were clonal and 9.5% subclonal but, when stratifying by APOBEC (SBS2 and SBS13) or AID mutations, the proportion was slightly higher within subclonal mutations (2.8% versus 1.6% in SBS2, 2.5% versus 1.3% in SBS13 & 5.2% versus 4.4% in AID; $P \sim 0$, respectively; fisher exact test; Supplementary Figure 17A). However, when looking only at samples with significant change within the mutational spectra of clonal versus subclonal mutations (404/2,707 samples; $P < 0.05$, Bonferroni-adjusted likelihood-ratio test), we found that even though AID-related mutations are slightly more

subclonal globally (median fold change = 1.06; $IQR = 0.85 - 1.31$), it has different temporal preference across tumor types. For example, there was a 9.6 fold change towards clonal mutations ($IQR = 3.9 - 33.8$) in Skin-Melanoma but a 1.3 fold change towards subclonality ($IQR = 0.83 - 1.61$) in Breast adenocarcinoma. On the other hand, we found more stable behaviors towards APOBEC (SBS2 and SBS13, more subclonal) or non-canonical AID (SBS9, more clonal) associated COSMIC signatures across tumor types and in the same direction, as previously described (Nicholas McGranahan et al., 2015; PCAWG Mutational Signatures Working Group et al., 2020) (Figure 5.5D). Furthermore, separating mutations by replication zones, globally and per tumor type, showed no evident change suggesting that this temporal preference is independent of the replication timing (Supplementary Figure 17B). Next, we checked if there were genes enriched for clonal or subclonal mutations globally or if AID provoked only, by calculating the odds ratio (OR) and corrected p-value of observed clonal/subclonal mutations versus the expected adjusting by various genetic covariates. Only a limited number of genes emerged as significantly enriched towards subclonality that were AID-induced (Supplementary Figure 17C). Summing up, we found that for some cancers like skin melanoma AID-mutations probably contribute to oncogenesis at early times meanwhile for others it might promote oncogenic fitness by late mutations.

AID synergizes initial hotspot mutations through late mutations on weakly functional alleles

Since recent studies have unraveled that composite mutations, pair of driver–driver, driver–passenger, or passenger–passenger mutations on the same gene, can synergize the functional impact compared to their single-mutated counterpart, we analyzed the contribution of AID induced mutations within this phenomenon by analyzing 31,353 samples comprising 41 tumor types from the MSKCC cohort (Figure 5.6). As previously described, using a panel of 353 oncogenes (168 genes) or tumor suppressor genes (TSGs, 185 genes), we found that composite mutations occur more frequently in TSGs than in oncogenes (12.2% versus 6.0% of all mutations; $P = 2e^{-278}$, two-sided two-sample Z-test) (Gorelick et al., 2020; Saito et al., 2020) but interestingly when separating by AID induced compared to those of other origins, we observed a global contribution to the composite mutations of 6.9%; furthermore, within oncogenes, 9% consisted of at least one AID induced mutation, compared to 5% within TSGs (Supplementary Figure 18A). We further verified that biallelic loss was also enriched for AID composite mutations, as it was reported from global composite mutations, within TSGs since there were more truncating variants compared to oncogenes (64% versus 8%; $P \sim 0$; Fisher exact test; Supplementary Figure 18B). Next, we calculated gene enrichment for

AID composite mutations globally and per tumor type to discard that the observations were due to randomness by modeling the AID composite mutational burden as a function of genetic covariates (see Methods). Surprisingly, we found enrichment for six genes including *FGFR3* especially among HNSCC with 20% corresponding to AID composite mutations, and lower lineage-specific proportions for *EGFR* (8.9% in Glioma), *PIK3CA* ($\sim$ 4% in Breast, Endometrial, Cervical, and Skin cancers), *FBXW7* ($\sim$ 7% in Colorectal and Esophagogastric cancers); *PTEN* (2.5 and 4% in Endometrial and Cervical cancers) but not *TP53* since it was present across different tumor types ($Q < 0.01$; Figure 5.6A and Supplementary Figure 18C, Supplementary Tables 4-5). We used a similar approach for residue's enrichment to avoid missing residues not enriched at the gene level and observed that *PIK3CA E726* was the most enriched ($q = 2.59e^{-58}$, Fisher's exact test) followed by *TP53 R213*, *EGFR A289*, and *PIK3CA R88* (Figure 5.6B, Supplementary Table 6). Since most found residues happened to be of lesser positive selection, we next checked the cumulative proportion moving from frequent hotspots (greatest positive selection) to less frequent ones finding that AID composite mutations are five times more likely to happen than AID singleton mutations ($P = 2e^{-109}$, two-sample Z-test for equal proportion) which has higher than the fold change (FC) between composite mutants (other than AID) to singleton mutants ($FC = 2.3$; $P \sim 0$). Furthermore, any AID mutation was absent from the highest positive selective hotspots (i.e. *KRAS G12*, *PIK3CA H1047*, *TP53 R273*) suggesting that AID mutations have a preference towards weakly functional alleles after the acquisition of high positive hotspots (Figure 5.6C, Supplementary Table 7). To further evaluate this hypothesis, we added the allelic configuration and clonality to subset to mutations arising from the same tumor cell population and retain molecular timing information. We observed that both globally (69% versus 31%, $P = 7e^{-4}$, two-sided binomial test, Supplementary Figure 18D) and within AID composites (73% versus 27%, $P = 0.03$, two-sided binomial test, Figure 17E) the most frequent hotspot mutation occurs first and is followed by a synergizing second mutation but only within oncogenes, which was the case of the minor mutation *PIK3CA E726*, between the kinase and the PI3KA domains, that occurs significantly after ($p = 0.039$, one-sided binomial test) than other stronger mutations (i.e. *PIK3CA E542*, *PIK3CA E545* at helical domain or *PIK3CA H1047* at the kinase domain) (Figure 5.6D, Supplementary Table 8) and is a product of AID promiscuous activity. When looking only at phase-able mutations (without the molecular timing variable) we observed that 88% of composite mutations on *PIK3CA* occur in cis from which 26% were AID provoked; other genes with a high percentage of cis AID composite mutations were EGFR, KMT2D, and APC (Supplementary Figure 18F, Supplementary Table 9). Some *PIK3CA* composite mutations have already been proved to increase cell proliferation, tumor growth

but also PI3K inhibitor sensitivity in human breast epithelial cell lines, but to the best of our knowledge, it has not been linked to being the product of AID activity (Saito et al., 2020; Vasan et al., 2019).

Additionally, we analyzed the contribution of other mutational processes to the composite mutations. Besides the aging signature, AID contributed more to the composite mutations than other signatures (Supplementary Figure 19), opening the possibility of further research on the molecular implications of these mutations.

The impact of AID-related mutations with ICI response

Because several recent studies pinpointed a potential role of APOBEC related mutations on the efficacy of ICI (Litchfield et al., 2021; S. Wang, Jia, He, & Liu, 2018), we sought to use the fraction of AID as a surrogate marker of ICI response. We used different available datasets analyzed (see Methods). We performed a random-effects meta-analysis comparing the overall survival (OS) of all these studies and comparing the impact of AID, to the APOBEC signature and the different single nucleotide variants (SNV). The details of this analysis are provided in the methods. Strikingly, the AID-related mutations were associated with the best OS in all of the studies and the random-effects model showed also a favorable prognosis (median as the cut-off, Figure 5.7A). Moreover, the effect was still significant across almost all the studies independently of either decile chosen as cut-off at univariate (Supplementary Figure 20A) or multivariate adjusting for TMB (Supplementary Figure 20B). Accordingly, the APOBEC signature was associated with a favorable prognosis, but not in all datasets. However, the random-effects model also indicated an overall favorable prognosis associated with APOBEC. The rest of SNV showed much more heterogeneous results and only T>A and T>G mutations were associated with favorable prognosis in the random-effects model (Figure 5.7A).

Interestingly, within the largest study of IMPACT-MSKCC, the fraction of AID-related mutations (top 50% of all histologies as the cut-off) was also independently associated with both better OS (Hazard ratio [HR] = 0.715; $95\% CI$ = 0.61-0.839; $p = 3.81 \times 10^{-5}$) and predictive value compared to TMB or APOBEC after adjusting by TMB (top 20% of each histology as the cut-off), APOBEC signature (top 50% of all histologies as the cut-off) age and sex (Figure 5.7B). It should be noted, that when using a univariate Cox proportional Hazards ratio model per every cancer type or adjusting by $TMB \geq 10$, the results were also similar in the overall population of this study, but the clinical impact of AID-related fraction of mutations was only found in metastatic melanoma and cancer with unknown primary (Figure 5.7C; Supplementary Figures 20C and 20D). Additionally, there was practically no correlation between the fraction of AID mutations with the APOBEC signature neither globally nor by tumor type in this cohort and in the ICGC and TCGA datasets (Supplementary Figures 21A-C). Similarly, by using

four additional studies across different tumor types, we also found an association of high AID mutations with improved OS after adjusting by age, gender, and TMB using the multivariate Cox model (Hugo et al., 2016; D. Liu et al., 2019; Miao et al., 2018; Pender et al., 2021).

Overall, all the studies confirmed the independent prognostic value of the high fraction of AID mutations according to the median in the univariate and multivariate analyses.

Landscape of AID-related neoepitopes and its relation with ICI response

Having found an association between AID activity and ICI benefit, we hypothesized that AID mutations might generate highly immunogenic neoepitopes. We addressed this by analyzing the neoepitopes that were products of AID activity on the TCGA cohort and on melanoma patients treated with Nivolumab (anti-PD-1) (Riaz et al., 2017). A recent bioinformatic-experimental study using immunogenic and non-immunogenic peptides, experimental testing, and X-ray structures showed that TCR binding and recognition improves with the presence of hydrophobic amino acids (aromatic W, F, Y followed by V, L, and I) at specific "MIA" positions (position $P_4 - P_{\Omega-1}$) due to increased structural avidity, stacking interactions, hydrogen bond acceptance and limited rotational freedom with the TCR (Schmidt et al., 2021). Additionally, as a previous study showed that APOBEC promiscuous activity increases neopeptide hydrophobicity (Boichard et al., 2018), we wondered if AID-related mutations led to the production of not only more hydrophobic neoepitope but more "Immunogenic" in terms of amino acid changes (W, F, Y, V, L, I over others) at MIA positions and if these effects were different due to clonality, histology, or mutational processes. We computed the PRIME %rank score and used it to classify neoepitope as "Immunogenic" or "Non-Immunogenic" (see Methods), on a list comprising 2,143 patients (TCGA) from which RNA-seq, HLA haplotyping, clonality, and mutational process origin data were correctly assessed; we also restricted the analysis to only patients with >1 FPKM expression on the genes originating the neopeptide, microsatellite stability, and intact antigen presentation related genes. We analyzed 286,909 neoepitopes from which only 17.75% were predicted to be immunogenic but interestingly they occur more frequently within clonal neoepitope than in subclonal (38% versus 30%, $P = 1 \times 10^{-27}$; two-sided Fisher-exact test; Figure 5.8A). Because our results suggested a higher presence of immunogenic neoepitope, in terms of numbers, provoked by mutations occurring earlier, we restricted the subsequent analyses to only ICNs.

Strikingly, albeit a global higher number of APOBEC induced ICNs was present, the proportion of samples having at least one ICN produced by AID (classified as "Present") was practically three times higher than those provoked by APOBEC globally (32% versus 11%, $P = 1 \times 10^{-65}$; two-sided Fisher-exact

test; Figure 5.8A). We next sought to compare the cumulative distribution of the AID/APOBEC ICNs in terms of population hotspot mutations recurrency finding that AID produces ICNs at hotspots with greater positive selection ($FC = 1.59$; $P = 3e^{-41}$, two-sample Z-test for equal proportion, Figure 5.8A) which could give rise to higher possibilities of immune recognition and improved tumor control. By comparing tumors harboring at least one AID ICN ("Presence") to those which did not ("Absence"), we found an increased fraction of CD8, CD4 memory activated and follicular helper T-cells that were "exhausted" by higher expression of the inhibitory immune checkpoint molecules PD-1, PD-L1, PD-L2, CTLA-4, and LAG3. Furthermore, these observations were seen in the majority of tumor types but the increment was only significant when accounting for all the samples (n = 2,143) or for LUAD (Figure 5.8B, two-sided Wilcoxon test).

Since these findings suggested that AID mutations inducing ICN as a possible explanation of ICI response, we next analyzed a cohort of 68 melanoma patients treated with anti-PD-1 (Nivolumab) from which WES, neoepitopes, and RNA-seq data were available prior treatment (pre) or 4 weeks after initiation of Nivo (on) (Riaz et al., 2017). Through all the analysis we separated patients as Ipi-Prog (n = 35), which had previously progressed on anti-CTLA-4 treatment (Ipilimumab), or as Ipi-Naive, which only received Nivo (n = 33). First, we looked at the distribution and effect of AID mutational load on survival compared to UV-related mutations. We found that the responders (CP/PR) had a higher number of AID-related mutations compared to the PD or SD groups (Ipi-Prog median = 0.094; Ipi-Naive median = 0.108), but was not significantly different and was also observed for UV mutations (Ipi-Prog median = 0.354; Ipi-Naive median = 0.349). Conversely, the effect on OS was markedly different, being associated with prognosis only when using AID mutations within Ipi-Naive patients (log-rank p = 0.026) but not with UV mutations in neither naive nor progressive patients (log-rank p = 0.93 & $p = 0.34$; Supplementary Figure 22A-B). The AID ICN load improved survival prediction better (log-rank p = 0.0016) than if using global clonal neo-epitopes load (log-rank p = 0.0042), global ICN load (log-rank p = 0.0025) or UV ICN load (log-rank p = 0.0071; Figure 5.8C). As the effect was tightly marked only in Ipi-Naive patients, we focused the subsequent analysis on only this group.

When coupling RNA-seq data (n = 20), we found 64 upregulated and 110 downregulated genes comparing patients with high AID ICN load versus low within pretherapy samples ($q < 0.20$; Supplementary Table 10). Gene Ontology (GO) analysis identified downregulation of antigen presentation and TNF signaling pathways (q-value < 0.05; Figure 5.8D & Supplementary Figure 22C). We also observed an increased expression of the inhibitory immune checkpoint molecules PD-1, PD-L1, PD-L2, CTLA-4, ICOS, LAG3, and cytolytic activity (Supplementary Figure 23).

These results are consistent with both our previous analysis on TCGA data and previous studies (N. McGranahan et al., 2016; Riaz et al., 2017; Van Allen et al., 2015).

Next, we endeavored to identify expression changes on patients that responded (according to AID ICN load) after 4 weeks of Nivo treatment by comparing pre-therapy to on-therapy data from the patients (npre = 20; non = 20). From the 811 genes found to be differentially expressed ($q < 0.20$; Supplementary Table 11), 404 were upregulated and involved in antigen processing and presentation, T-cell activation (e.g. *PRKCQ*, *CD8B*, *CD38*, *CD151*, *MALT1*), leukocyte cell-cell adhesion, response to oxidative stress (*STK24*, *GSS*, *GCLC*, *PDK1*) and T-cell reactivity to clonal neoepitopes (*CXCL13* and *CCR5*) (q-value < 0.05), the last ones being recently described (Litchfield et al., 2021). On the other hand, down-regulated pathways included mainly (407 genes; q-value < 0.05) cell growth, B-cell differentiation, and some chemokines (*CXCL11*, *CCL4*, and *CCL14*) or chemokine receptors (*CCR3* and *CCR8*) (Figure 5.8D). Furthermore, we also observed an increased expression of *CXCL13* and *CCR5* on the TCGA samples with high AID ICN load (Figure 5.8B). Altogether, these results show possible explanations of why AID mutations reflect a more straightforward approach to predict response to ICI naive treatment.

5.2.3 Discussion

By integrating more than 50.000 bulk level samples and 2.5 million cells at single-cell resolution across 80 tumor types and different data levels, we present, to the best of our knowledge, the first study shedding light on the oncogenic and clinical implications of AID at pan-cancer scale. Our results point to the idea that *AICDA* expression, which is activated after malignant transformation, is no longer tied to the cell cycle regulation and, albeit transient, it induces traceable mutations with important functional and clinical implications that are mainly produced during the transcriptional activity of the mutated gene. Firstly, our single-cell RNA-seq analysis revealed that only a selected number of tissues (e.g. adult pleura, small intestine) express *AICDA* under normal conditions, which is consistent with previous studies at bulk level (Lonsdale et al., 2013; Uhlen et al., 2015); however, after acquiring malignancy the expression is seen independently of tissue origin as already reported for some tumor types (Endo et al., 2007; Kasar et al., 2015; Komori et al., 2008; L. Li et al., 2019; Lossos, Levy, & Alizadeh, 2004; Matsumoto et al., 2007; Nonaka et al., 2016; Sawai et al., 2015; Shimizu et al., 2014) . Moreover, our single-cell data analysis shows that B-cells within the TME of some cancer types are expressing AICDA which correlates with our findings on the TCGA data showing correlations of *AICDA* expression with the B-cell and lymphocyte infiltrate populations.

It was previously reported that in germinal center B-cells the *AICDA* expression is higher at G2/M cell cycle phases but its mutational effects are mostly active in early G1 (after being transported to the nucleus). However, this regulation is lost in lymphoma cells (Milpied et al., 2018; Q. Wang et al., 2017). On the other hand, UNG expression is most abundant at the S phase meanwhile MSH2/MSH6 at G1/S, both implicated in repairment of AID-induced mutations (Álvarez-Prado et al., 2018; Delgado et al., 2020; Kasar et al., 2015). The oncogenic single-cell datasets analyzed proved that this cell cycle regulation is globally lost in all tumor types and that cells at G2/M phases are more susceptible to AID promiscuous activity given that: i) *AICDA* expression is higher; ii) the BER and MRR related genes are less expressed and iii) global transcriptional activity is also increased. This hypothesis is supported by our findings, using WGS data, that AID mutational load is increased: at transcriptionally active TAD domains (compared to the background), close to TSS, and in MSI tumors. Regarding the different AID mutational behavior depending on the strand location of the gene, we propose a model where the negative strand is more prone, than the positive strand, to AID attack at naked transcribed breathing dsDNA (normally located near TSS) and is followed by attack at DNA stem-loops and transcription bubbles (but not at R-loops) being generated as the RNA polymerase transcribes (Branton et al., 2020).

Summing up to those findings plus that *AICDA* expression and AID related mutations were not correlated in the TCGA nor in ICGC datasets and that only expression but not the mutations correlated with viral infection in some cancers, it is tempting to speculate that the genotoxic effect of AID might be due to short term activation of *AICDA*, which have been seen in APOBEC (Langenbucher et al., 2021). Indeed, in a fate-mapping study, *AICDA* expression was present in a fraction of non-lymphoid embryonic cells (Rommel et al., 2013). Furthermore, *AICDA* transcripts in lymphocytes have a half-life of only one hour (Dorsett et al., 2008), supporting the lack of correlation between AID-related mutations and *AICDA* expression.

Despite, ephemeral, *AICDA* expression mutational footprints are widespread across cancers, and presumptively across mammals, with similar mutational frequency compared to APOBEC but a higher contribution to driver oncogenes, to composite mutations, and to the production of higher quality neo-epitopes. Already reported AID off-target activity, outside lymphomas, is limited especially to *TP53*, *KRAS*, and *MYC* in gastric, colorecta,l and skin melanoma (Hanjie Li et al., 2019; Nonaka et al., 2016; Shimizu et al., 2014). We thoroughly extended this data and found that AID activity has a preference towards least positive selection hotspots that synergizes with previous stronger hotspot mutations; this is the case for the minor mutation PIK3CA E726, especially present in SKCM and BRCA,

that might confer higher PI3K inhibitor sensitivity (Saito et al., 2020; Vasan et al., 2019).

Finally, we found that the AID-related fraction of mutations is an independent prognostic value to ICI response using $>$ 2,000 samples even after adjusting by TMB. AID-related neoepitopes exhibited distribution towards clonal hotspots with a greater positive selection which could result in improved immune recognition; however, this is avoided by tumor-induced immune exhaustion. It should be noted that the statistical power in individual histologies is reduced, and as sample sizes increase, additional histology-specific associations may appear in future larger prospective studies that may lead to a formal validation of the predictive value of AID-related signature on ICI response and the results regarding the AID-related neoepitopes. It is also important to highlight that there could be some analytical bias related to the combination of different datasets using different mutation calling approaches. However, the signal associated with the AID-related mutations was similar throughout the studies and the pipelines, and results of the different included studies are public and well standardized, limiting in part this mutation call bias.

We propose a model in which AID ICN has higher probabilities of being recognized by T-cells, triggering selective expression *CXCL13*, previously found to be a marker of antigen reactive CD8 T-cells, for recruitment of CXCR5+ T and B cells (Litchfield et al., 2021). These recruited cells, subsequently exhausted by the continuous expression of inhibitory immune checkpoint molecules, can be reinvigorated after ICI treatment.

Overall, we pieced together an immense part of the oncogenic AID puzzle but many parts still need to be found, especially filling gaps with biological validations as the results, here presented, hold the promise of important clinical applications.

5.2.4 Acknowledgments

We greatly thank all investigators, funders, and industry partners that supported the generation of the data within this study, as well as patients for their participation. Specifically, we thank Eli Van Allen, Roel Verhaak, and Timothy A Chan for the academic datasets. The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga and the ICGC consortium: https://dcc.icgc.org. Graphical abstract created with biorender.com.

5.2.5 Funding

This work was in part supported by a grant from Investissements d'avenir and by the grant INCa-DGOS-Inserm_12560 of the SiRIC CURAMUS, the program

"investissements d'avenir" ANR-10-IAIHU-06, PRT-K/INC a grant LOC-model reference 2017-1-RT-04, BETPSY project, overseen by the French National Research Agency, as part of the second "Investissements d'Avenir' ' program (Grant No. ANR-18-RHUS-0012), RAM foundation, ARTC foundation, an unrestricted grant from Bristol Myers Squibb (BMS): RDON06618 and IDeATIon project with an unrestricted grant from MSD Avenir. D.R. was partially supported by a Bicocca 2020 Starting Grant and by a Premio Giovani Talenti dell'Università degli Studi di Milano-Bicocca.

The funders had no roles in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

5.2.6 Author contributions

I.H.-V. and A.A. designed the study. I.H.-V., K.M., K.H.-X.,M.P., F.B., M.T., A.I., A.D.,M.S. and A.A. performed clinical work. I.H.-V., K.C.A., D.R., G.C., K.L. and A.A. analyzed data. I.H.-V., K.C.A.,D.R.,G.C., A.D., M.S. and A.A. interpreted data. I.H.-V. and A.A. wrote the manuscript.

5.2.7 Declaration of interest

Dr. Ahmed Idbaih reports grants and travel funding from Carthera, research grants from Transgene/ Sanofi/Air Liquide/Nutritheragène, advisory board personal fees from Novocure/LeoPharma, outside the submitted work. Dr. Franck Bielle reports interests out of the scope of this article: (i) a next-of-kin employed by Bristol Myers Squibb, (ii) funding of research by Abbvie, (iii) fees of travel and conference funded by Bristol Myers Squibb.

5.2.8 STAR⋆METHODS

Methods (including statements of data availability and any associated accession codes and references) and supplementary material are available in the online version of the paper.

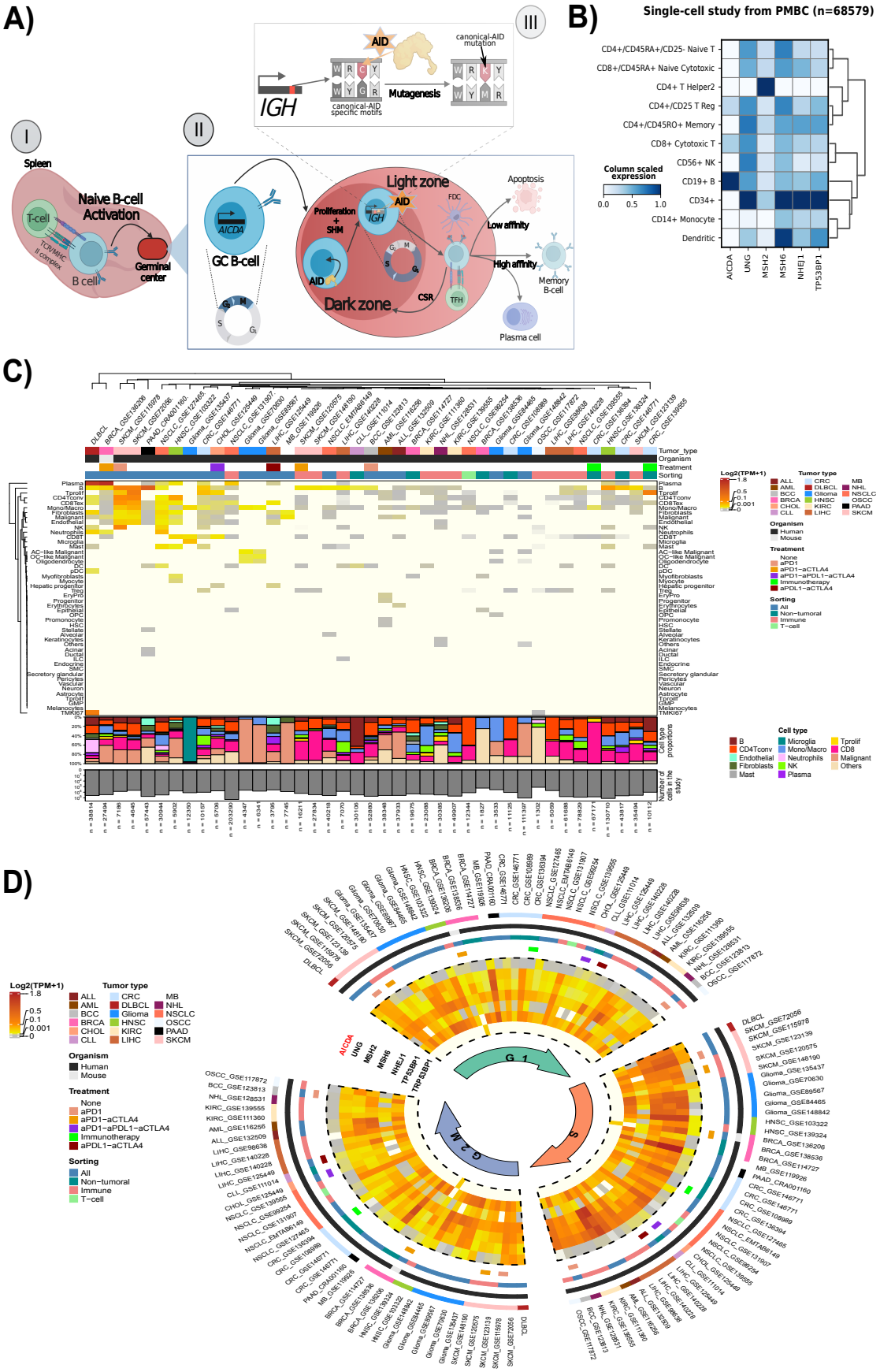
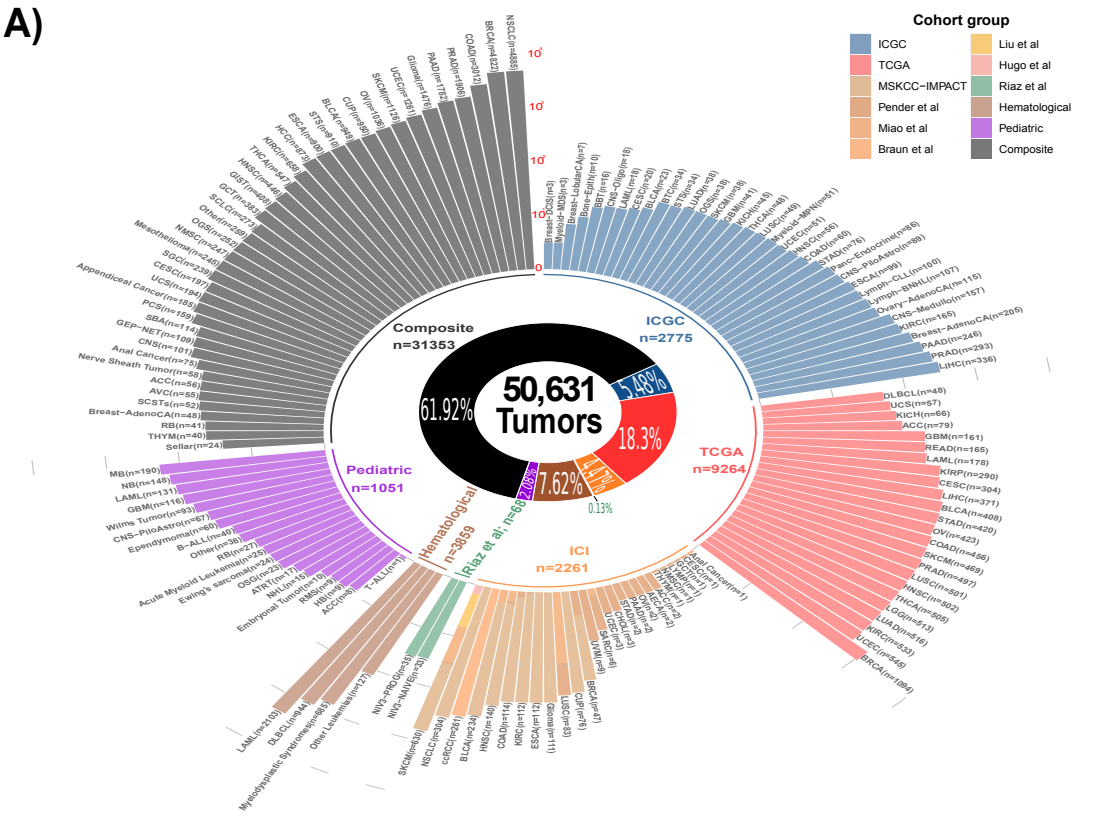
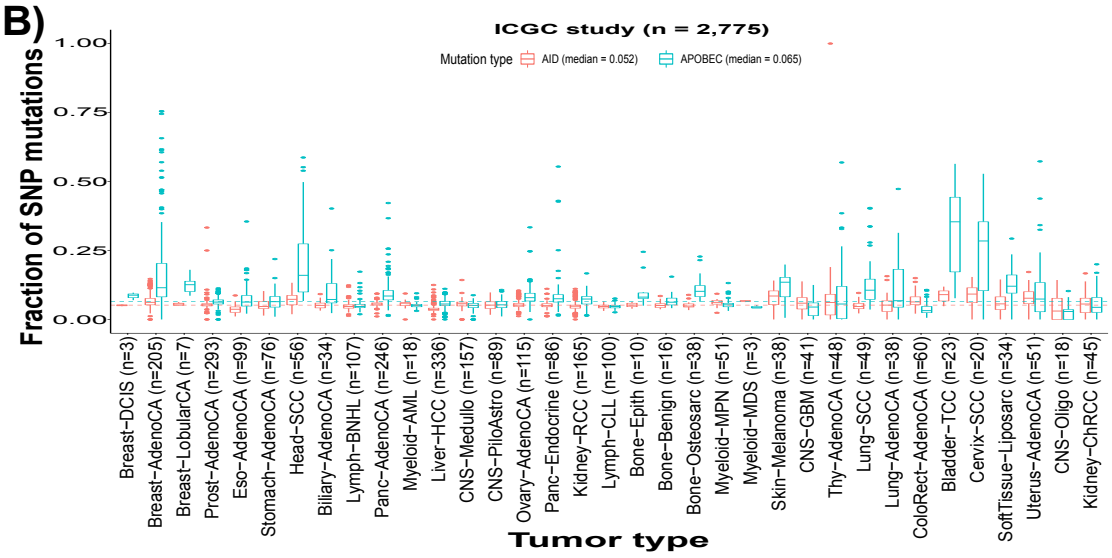

A)
III
AID
canonical-AID mutation
IGH
Mutagenesis
canonical-AID specific motifs
I
II
Spleen
Naive B-cell Activation
T-cell
TCR/MHC II complex
B cell
Germinal center
AICDA
GC B-cell
Light zone
Proliferation + SHM
AID
IGH
FDC
Apoptosis
Low affinity
High affinity
Memory B-cell
Dark zone
CSR
TFH
Plasma cell
B)
Single-cell study from PMBC (n=68579)
CD4+/CD45RA+/CD25- Naive T
CD8+/CD45RA+ Naive Cytotoxic
CD4+ T Helper2
CD4+/CD25 T Reg
CD4+/CD45RO+ Memory
CD8+ Cytotoxic T
CD56+ NK
CD19+ B
CD34+
CD14+ Monocyte
Dendritic
Column scaled expression
0.0 0.5 1.0
AICDA
UNG
MSH2
MSH6
NHEJ1
TP53BP1
C)
Tumor_type
Organism
Treatment
Sorting
Cell type proportions
Number of cells in the study
Log2(TPM+1)
Tumor type
Organism
Treatment
Sorting
Cell type
D)
Log2(TPM+1)
Tumor type
Organism
Treatment
Sorting
G 1
S
G 2 M

Figure 5.2 *(preceding page)*: scRNA analysis reveals activation of *AICDA* expression under oncogenic conditions. Panel A, illustrative representation of *AICDA* expression and AID activity within normal B-cells. For AID motifs, W = A/T; R = Purine; Y = Pyrimidine; K = G/T; M = C/A. Panel B, normal PBMCs expression heatmap showing *AICDA* expression only in B-cells, CD8 T-cells, and induced regulatory T-cells. Panel C, *AICDA* expression heatmap as a function of the cell subtypes across different oncogenic single-cell studies (n = 41; top), barplot of cell subtypes proportions (middle), and the number of cells in each study (bottom). *AICDA* is expressed in malignant cells in most tumor types but only in some for the immune and stromal cell populations. Panel D, expression heatmaps of *AICDA* and genes involved in the repair of *AICDA*-related mutations as a function of the cell cycle stage across different oncogenic single-cell studies (n = 40). *AICDA* expression is slightly higher in the G2M phase and the repair genes are more expressed during the S phase; additionally, expression is dropped when tumoral cells are depleted by sorting.


A)
Cohort group
ICGC
TCGA
MSKCC−IMPACT
Pender et al
Miao et al
Braun et al
Liu et al
Hugo et al
Riaz et al
Hematological
Pediatric
Composite
50,631 Tumors
Composite n=31353
ICGC n=2775
TCGA n=9264
Pediatric n=1051
Hematological n=3859
ICI n=2261
Riaz et al.; n=69
61.92%
5.46%
18.3%
7.62%
2.08%
4.17%
0.13%


B)
ICGC study (n = 2,775)
Mutation type
AID (median = 0.052)
APOBEC (median = 0.065)
Fraction of SNP mutations
1.00
0.75
0.50
0.25
0.00
Tumor type

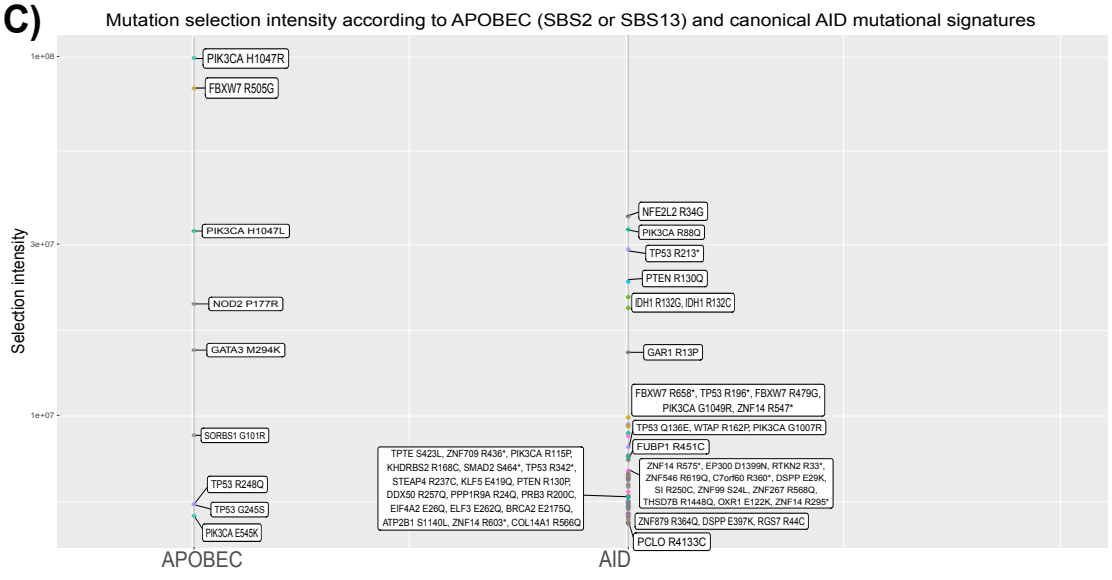

C)
Mutation selection intensity according to APOBEC (SBS2 or SBS13) and canonical AID mutational signatures
Selection intensity
1e+08
3e+07
1e+07
PIK3CA H1047R
FBXW7 R505G
PIK3CA H1047L
NOD2 P177R
GATA3 M294K
SORBS1 G101R
TP53 R248Q
TP53 G245S
PIK3CA E545K
NFE2L2 R34G
PIK3CA R88Q
TP53 R213*
PTEN R130Q
IDH1 R132G, IDH1 R132C
GAR1 R13P
FBXW7 R658*, TP53 R196*, FBXW7 R479G, PIK3CA G1049R, ZNF14 R547*
TP53 Q136E, WTAP R162P, PIK3CA G1007R
FUBP1 R451C
ZNF14 R575*, EP300 D1399N, RTKN2 R33*, ZNF546 R619Q, C7orf60 R360*, DSPP E29K, SI R250C, ZNF99 S24L, ZNF267 R568Q, THSD7B R1448Q, OXR1 E122K, ZNF14 R295*
ZNF879 R364Q, DSPP E397K, RGS7 R44C
PCLO R4133C
TPTE S423L, ZNF709 R436*, PIK3CA R115P, KHDRBS2 R168C, SMAD2 S464*, TP53 R342*, STEAP4 R237C, KLF5 E419Q, PTEN R130P, DDX50 R257Q, PPP1R9A R24Q, PRB3 R200C, EIF4A2 E26Q, ELF3 E262Q, BRCA2 E2175Q, ATP2B1 S1140L, ZNF14 R603*, COL14A1 R566Q
APOBEC
AID

Figure 5.3 *(preceding page)*: Pan-cancer landscape of AID-related mutations. Panel A, samples' distribution of the different cohorts used in the study (51,631 tumors) where each bar is a tumor type or subgroup (only Riaz et al cohort); for the ICI cohort stacked bars represent the different studies. The complete 88 tumor types' abbreviations are presented in Supplementary Table 2. Panel B, frequency of the fraction of mutations attributed to AID motifs or APOBEC motifs for each tumor type in the ICGC cohort; dotted lines indicate median values across all samples. Panel C, AID mutations produce higher selection intensity on driver genes on minor hotspot residues but there is a higher number of affected genes/residues than the ones generated by the APOBEC related signatures.

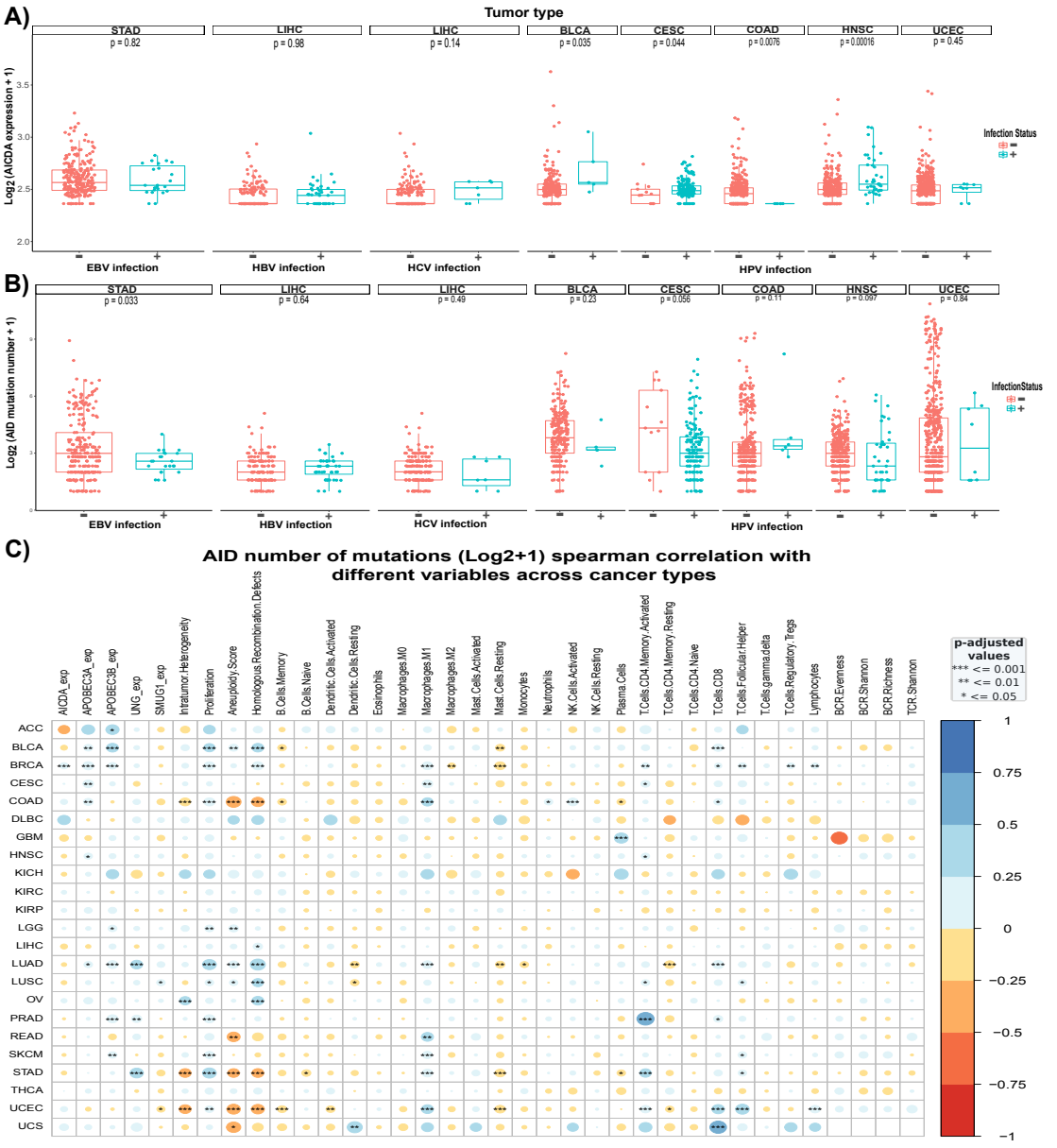

A)
Tumor type
STAD
LIHC
LIHC
BLCA
CESC
COAD
HNSC
UCEC
p = 0.82
p = 0.98
p = 0.14
p = 0.035
p = 0.044
p = 0.0076
p = 0.00016
p = 0.45
Log2 (AICDA expression + 1)
Infection Status
EBV infection
HBV infection
HCV infection
HPV infection
B)
p = 0.033
p = 0.64
p = 0.49
p = 0.23
p = 0.056
p = 0.11
p = 0.697
p = 0.84
Log2 (AID mutation number + 1)
InfectionStatus
C)
AID number of mutations (Log2+1) spearman correlation with different variables across cancer types
AICDA_exp
APOBEC3A_exp
APOBEC3B_exp
UNG_exp
SMUG1_exp
Intratumor.Heterogeneity
Proliferation
Aneuploidy.Score
Homologous.Recombination.Defects
B.Cells.Memory
B.Cells.Naive
Dendritic.Cells.Activated
Dendritic.Cells.Resting
Eosinophils
Macrophages.M0
Macrophages.M1
Macrophages.M2
Mast.Cells.Activated
Mast.Cells.Resting
Monocytes
Neutrophils
NK.Cells.Activated
NK.Cells.Resting
Plasma.Cells
T.Cells.CD4.Memory.Activated
T.Cells.CD4.Memory.Resting
T.Cells.CD4.Naive
T.Cells.CD8
T.Cells.Follicular.Helper
T.Cells.gamma.delta
T.Cells.Regulatory.Tregs
Lymphocytes
BCR.Evenness
BCR.Shannon
BCR.Richness
TCR.Shannon
p-adjusted values
*** <= 0.001
** <= 0.01
* <= 0.05
ACC
BLCA
BRCA
CESC
COAD
DLBC
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PRAD
READ
SKCM
STAD
THCA
UCEC
UCS
1
0.75
0.5
0.25
0
−0.25
−0.5
−0.75
−1

Figure 5.4 *(preceding page)*: AID-mutations and *AICDA* expression relation with immune features. Panel A, distribution of the expression of *AICDA* according to the presence of different viral genomes per cancer type. Panel B, distribution of the number of AID-related mutations according to the presence of viral genomes per cancer type. Panel C, heatmap showing the correlation according to the number of AID-related mutations and the expression of different immune cell types, genes of interest, or the presence of different viral genomes; circle size/color indicates the direction of association (Spearman correlation) and the annotation inside the circles indicate the significance (FDR-adjusted p-values).

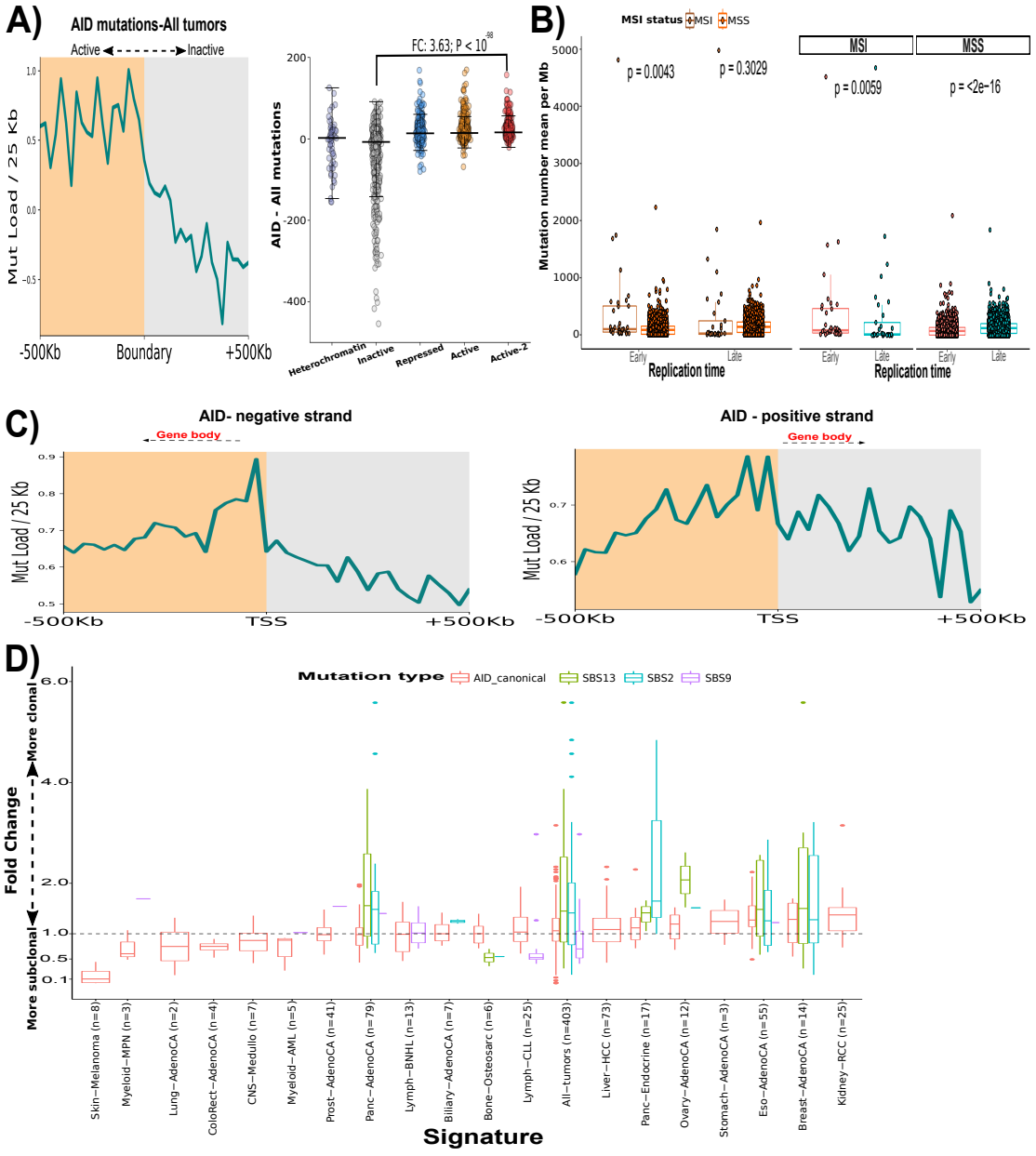

A)
AID mutations-All tumors
Active
Inactive
Mut Load / 25 Kb
-500Kb
Boundary
+500Kb
AID - All mutations
FC: 3.63; P < 10^-88
Heterochromatin
Inactive
Repressed
Active
Active-2
B)
MSI status MSI MSS
Mutation number mean per Mb
p = 0.0043
p = 0.3029
p = 0.0059
p = <2e-16
MSI
MSS
Early
Late
Replication time
C)
AID- negative strand
AID - positive strand
Gene body
Mut Load / 25 Kb
-500Kb
TSS
+500Kb
D)
Mutation type AID_canonical SBS13 SBS2 SBS9
Fold Change
More clonal
More subclonal
Skin−Melanoma (n=8)
Myeloid−MPN (n=3)
Lung−AdenoCA (n=2)
ColoRect−AdenoCA (n=4)
CNS−Medullo (n=7)
Myeloid−AML (n=5)
Prost−AdenoCA (n=41)
Panc−AdenoCA (n=79)
Lymph−BNHL (n=13)
Biliary−AdenoCA (n=7)
Bone−Osteosarc (n=6)
Lymph−CLL (n=25)
All−tumors (n=403)
Liver−HCC (n=73)
Panc−Endocrine (n=17)
Ovary−AdenoCA (n=12)
Stomach−AdenoCA (n=3)
Eso−AdenoCA (n=55)
Breast−AdenoCA (n=14)
Kidney−RCC (n=25)
Signature

Figure 5.5 *(preceding page)*: AID-mutations interplay with replication, transcription, and clonality. Panel A, average profile of AID somatic mutations accumulation in 2,775 cancer samples and replication timing across 500 kb of TAD boundaries delineating active to inactive domains (left); dot plots representing the distribution of the mutations in different domain-types (right; Wilcoxon rank-sum test). Panel B, distribution of the number of mutations associated with *AICDA* according to the replication time when considering MSS or MSI samples (panels at right). Panel C, average profiles of AID induced mutations accumulation in 2,775 cancer samples across 500 kb of TSS for negative-strand genes (left) or positive-strand genes (right). Panel D, fold change of signature activities (subclonal to clonal) across tumor types for AID, APOBEC (SBS2 and SBS13), and SBS9 mutational signatures for samples with measurables changes in their mutation spectra (n = 404) where box plots indicate the first and third quartiles of the distribution, with the median shown in the center and whiskers covering data within 1.5x the IQR from the box. Panel H, volcano plot showing the clonality enrichment per gene (FDR-adjusted from one-sided Fisher's exact test) against odds ratio where color indicates enrichment for AID or global somatic mutations. All panels were produced using the ICGC cohort.

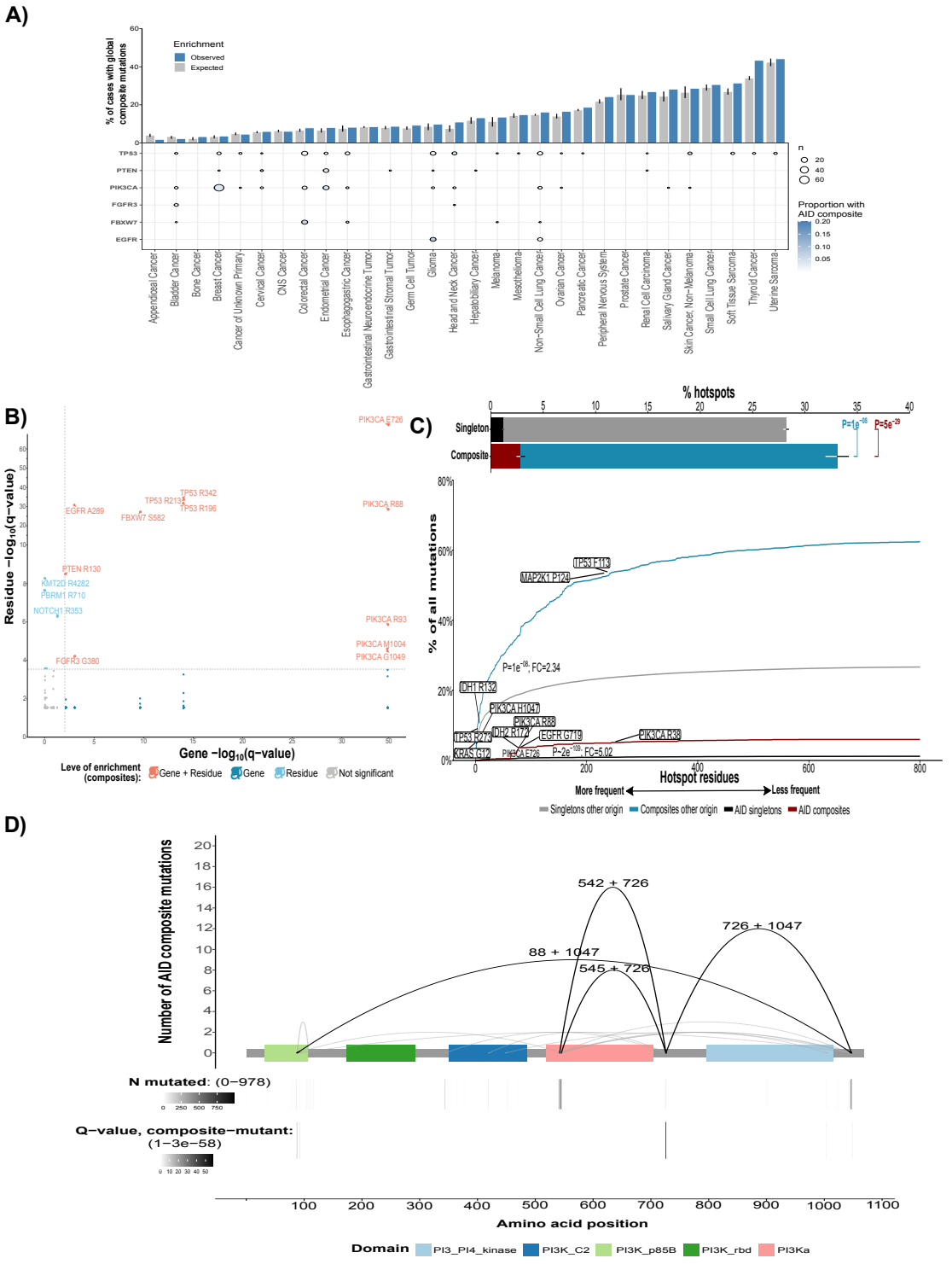

A)
% of cases with global composite mutations
Enrichment
Observed
Expected
TP53
PTEN
PIK3CA
FGFR3
FBXW7
EGFR
n
Proportion with AID composite
Appendiceal Cancer
Bladder Cancer
Bone Cancer
Breast Cancer
Cancer of Unknown Primary
Cervical Cancer
CNS Cancer
Colorectal Cancer
Endometrial Cancer
Esophagogastric Cancer
Gastrointestinal Neuroendocrine Tumor
Gastrointestinal Stromal Tumor
Germ Cell Tumor
Glioma
Head and Neck Cancer
Hepatobiliary Cancer
Melanoma
Mesothelioma
Non-Small Cell Lung Cancer
Ovarian Cancer
Pancreatic Cancer
Peripheral Nervous System
Prostate Cancer
Renal Cell Carcinoma
Salivary Gland Cancer
Skin Cancer, Non-Melanoma
Small Cell Lung Cancer
Soft Tissue Sarcoma
Thyroid Cancer
Uterine Sarcoma
B)
Residue −log10(q−value)
Gene −log10(q−value)
EGFR A289
TP53 R342
TP53 R213
TP53 R196
FBXW7 S582
PTEN R130
KMT2D R4282
PBRM1 R710
NOTCH1 R353
FGFR3 G380
PIK3CA E726
PIK3CA R88
PIK3CA R93
PIK3CA M1004
PIK3CA G1049
Leve of enrichment (composites):
Gene + Residue
Gene
Residue
Not significant
C)
% hotspots
Singleton
Composite
P=1e−16
P=5e−29
% of all mutations
TP53 F113
MAP2K1 P124
P=1e−08; FC=2.34
IDH1 R132
PIK3CA H1047
PIK3CA R88
IDH2 R172
EGFR G719
PIK3CA R38
TP53 R273
KRAS G12
PIK3CA E726
P=2e−169; FC=5.02
Hotspot residues
More frequent
Less frequent
Singletons other origin
Composites other origin
AID singletons
AID composites
D)
Number of AID composite mutations
542 + 726
726 + 1047
88 + 1047
545 + 726
N mutated: (0−978)
Q−value, composite−mutant: (1−3e−58)
Amino acid position
Domain
PI3_PI4_kinase
PI3K_C2
PI3K_p85B
PI3K_rbd
PI3Ka

Figure 5.6 *(preceding page)*: Impact of AID mutations on composite mutations. Panel A, AID composite mutations in enriched genes by lineage (n = 31,353 samples). Cases with global composite mutations and the expected value based on cohort size and mutational burden (top). Significant enrichment for AID composite mutations in cancer genes per cancer type (FDR-adjusted P values from a one-sided binomial test for enrichment; bottom, n = 29,461). Panel B, residue versus gene enrichment arising from AID composite mutations (FDR-adjusted from one-sided Fisher's exact test for residues or one-sided binomial test for genes). Panel C, cumulative sum of the percentage of hotspot mutation utilization by decreasing frequency of population-level hotspot mutations among composite or single mutations (AID or not AID provoked). Two-sided Mann–Whitney U test, fold-change (FC) of max composite to singleton values. Top inset, percentage of hotspots attributable to composite/singleton mutations (Two-sided two-sample Ztest for equal proportions, color indicates comparison for AID or not AID provoked); Error bars indicate 95% binomial confidence intervals (CIs). Panel I, occurrence of *PIK3CA* AID composite mutations where arcing lines indicate the composite pairs ($\geq$ 2 tumors, black color for AID enriched residues) and numbers indicate the amino acid position. Residue *PIK3CA E726*, located between the kinase and PI3KA domains, is highly enriched as an AID composite. Significance values for the composite mutants (FDR-adjusted P-value, one-sided binomial test) are shown at the bottom.

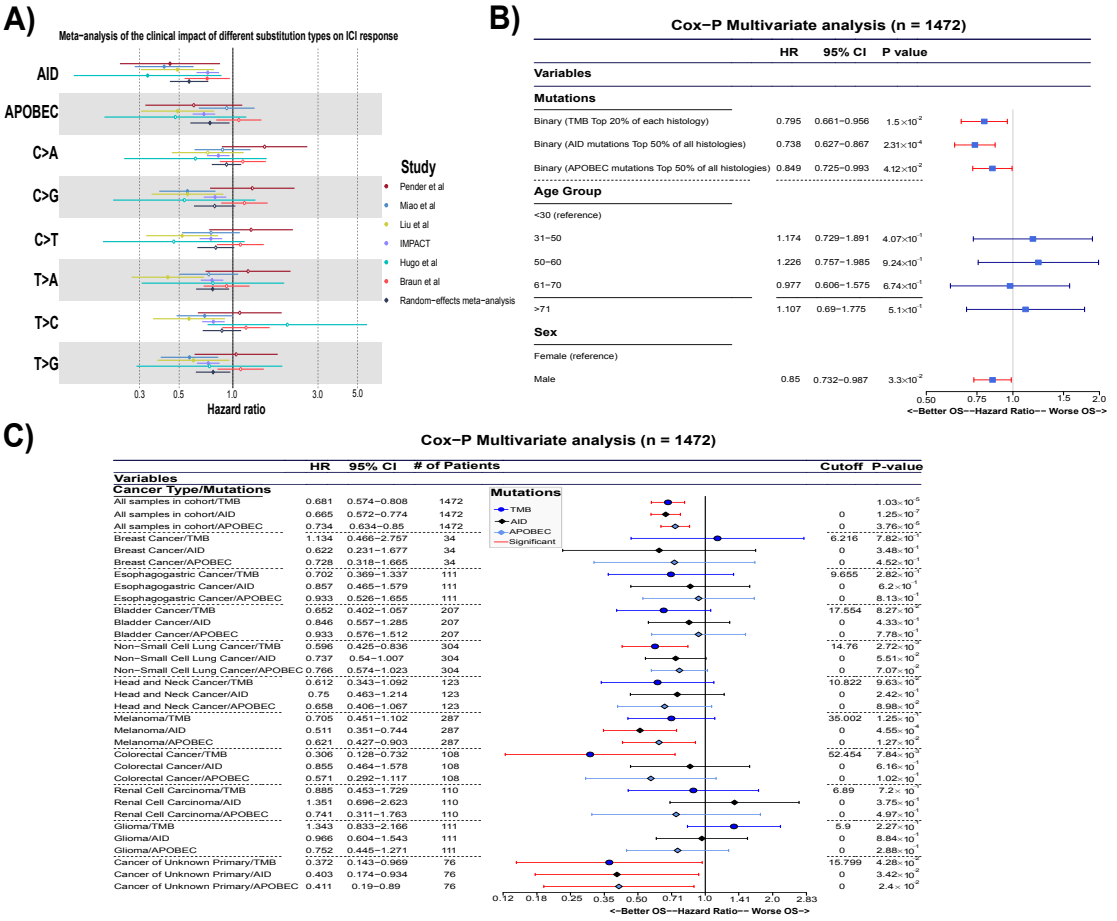


Figure 5.7: The impact of AID mutations on ICI response. Meta-analysis of the survival impact of the fraction of AID mutations in different studies. Panel A, effect of using AID/APOBEC (5th decile as cut-off) or SNV substitutions where AID remains significant across all the studies. Panel C, forest plot of a Cox model of the global impact, after adjustment by TMB (top 20%), median APOBEC mutations, age, and gender. Panel D, forest plot of the Cox model of the impact of AID mutations per cancer subtype.

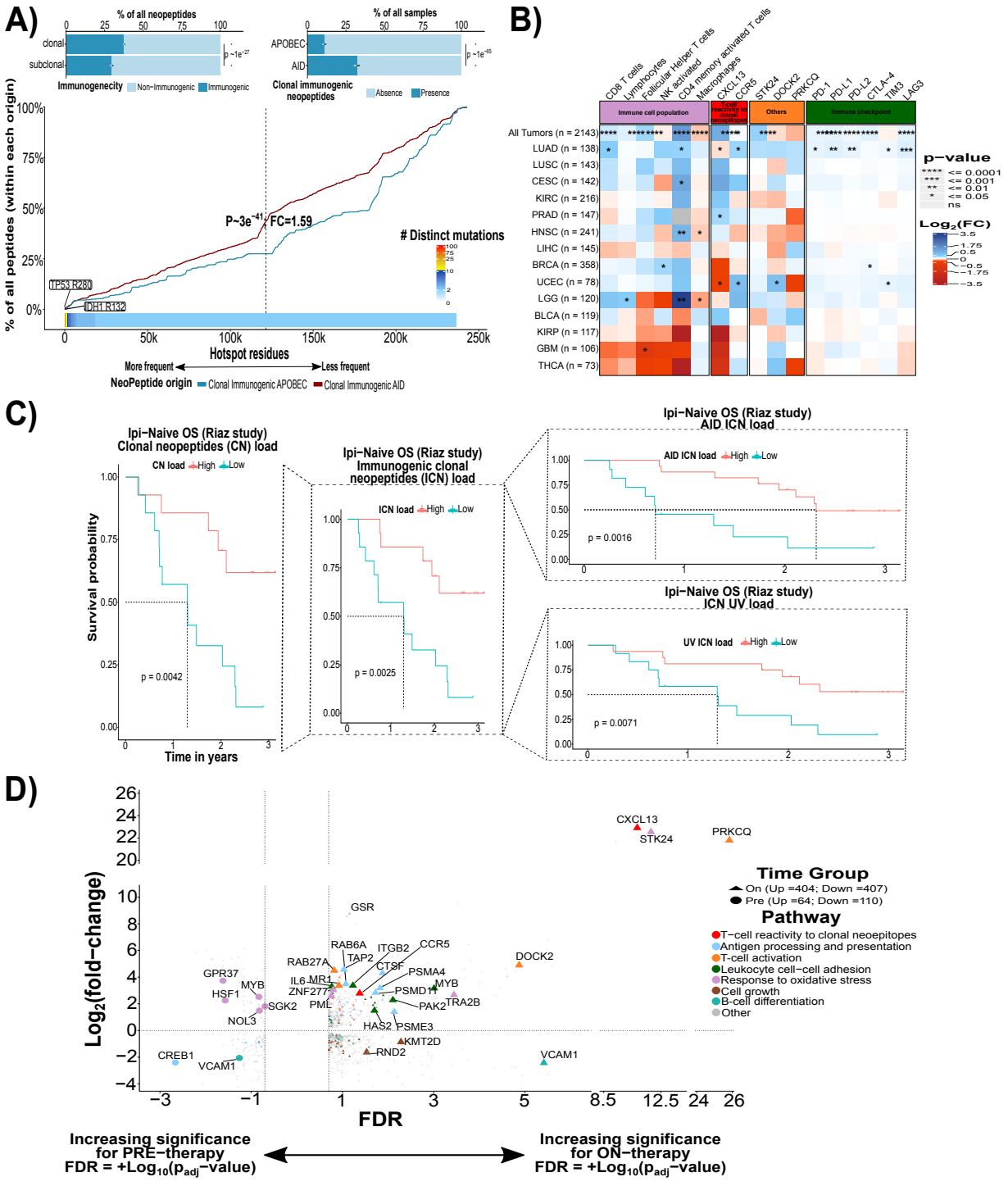

A)
% of all neopeptides
clonal
subclonal
Immunogenecity
Non−Immunogenic
Immunogenic
% of all samples
APOBEC
AID
Clonal immunogenic neopeptides
Absence
Presence
% of all peptides (within each origin)
P~3e−41 / FC=1.59
Distinct mutations
TP53 R280
IDH1 R132
Hotspot residues
More frequent
Less frequent
NeoPeptide origin
Clonal Immunogenic APOBEC
Clonal Immunogenic AID
B)
CD8 T cells
Lymphocytes
Follicular Helper T cells
NK activated
CD4 memory activated T cells
Macrophages
CXCL13
CCR5
STK24
DOCK2
PRKCQ
PD-1
PD-L1
PD-L2
CTLA-4
TIM3
LAG3
Immune cell population
Others
Immune checkpoint
All Tumors (n = 2143)
LUAD (n = 138)
LUSC (n = 143)
CESC (n = 142)
KIRC (n = 216)
PRAD (n = 147)
HNSC (n = 241)
LIHC (n = 145)
BRCA (n = 358)
UCEC (n = 78)
LGG (n = 120)
BLCA (n = 119)
KIRP (n = 117)
GBM (n = 106)
THCA (n = 73)
p−value
Log2(FC)
C)
Ipi−Naive OS (Riaz study)
Clonal neopeptides (CN) load
CN load High Low
p = 0.0042
Survival probability
Time in years
Ipi−Naive OS (Riaz study)
Immunogenic clonal neopeptides (ICN) load
ICN load High Low
p = 0.0025
Ipi−Naive OS (Riaz study)
AID ICN load
AID ICN load High Low
p = 0.0016
Ipi−Naive OS (Riaz study)
ICN UV load
UV ICN load High Low
p = 0.0071
D)
Log2(fold−change)
FDR
CXCL13
STK24
PRKCQ
GSR
RAB6A
ITGB2
CCR5
TAP2
RAB27A
CTSF
PSMA4
DOCK2
GPR37
MYB
IL6−MR1
PSMD11
MYB
HSF1
ZNF277
TRA2B
SGK2
PML
PAK2
NOL3
HAS2
PSME3
KMT2D
RND2
CREB1
VCAM1
VCAM1
Time Group
On (Up =404; Down =407)
Pre (Up =64; Down =110)
Pathway
T−cell reactivity to clonal neoepitopes
Antigen processing and presentation
T-cell activation
Leukocyte cell−cell adhesion
Response to oxidative stress
Cell growth
B-cell differentiation
Other
Increasing significance for PRE−therapy FDR = +Log10(padj−value)
Increasing significance for ON−therapy FDR = +Log10(padj−value)

Figure 5.8 *(preceding page)*: Landscape of AID-related neoepitopes and its relation with ICI response. Panel A, percentage of neoepitopes originating from clonal/subclonal mutations in which color indicates comparison for immunogenic or non-immunogenic (top left; Two-sided two-sample Ztest for equal proportions), calculated by Prime. Top right shows the comparison of the percentage of samples having at least one AICDA ICN versus APOBEC ICN ("Presence"; Two-sided two-sample Ztest for equal proportions). Bottom plot shows the cumulative sum of hotspot mutation utilization that gives rise to ICN by decreasing the frequency of population-level hotspot mutations due to AID or APOBEC signatures (Two-sided Mann–Whitney U test, FC of median AID to APOBEC values). Panel B, heatmap of gene expression comparison between AID ICN "Presence" versus "Absence" groups across tumor types/all tumors (n = 2,143; two-sided Wilcoxon test) measured as log2 FC. Panel C, OS prediction within Ipi-Naive patients improves when using AID ICN load (top right), ICN UV load (bottom right), ICN load (middle), or clonal neoepitopes load (left), lowest to highest log-rank p-values. Panel D, DEGs (p-adj <0.20) between high ICN load patients versus low ICN load for pre-therapy, where increasing negative values on the x-axis shows higher significance (+Log10[p-adj]), or on-therapy, where increasing positive values on the x-axis means higher significance (-Log10[p-adj]). The y-axis shows upregulated (FC > 0) or downregulated genes (FC < 0) and colors indicate genes enriched in a specific pathway by GO analysis. Panels A and B correspond to the TCGA cohort (n = 2,143) meanwhile panels C and D to ICI treated melanoma cohort (Riaz et al., n = 68).

5.3 Molecular and clinical diversity in primary central nervous system lymphoma

Isaias Hernández-Verdin[1], Eva Kirasic[1], Kirsty Wienand[2], Karima Mokhtari[1,3], Sandrine Eimer[4], Hugues Loiseau M.D. Ph.D., Audrey Rousseau M.D., Ph.D., Jérôme Paillassa M.D., Guido Ahle M.D., Felix Lerintiu M.D, Emmanuelle Uro-Coste M.D., Ph.D., Lucie Oberic M.D., Dominique Figarella-Branger M.D. Ph.D., Olivier Chinot M.D., Ph.D., Guillaume Gauchotte M.D. Ph.D., Luc Taillandier M.D., Ph.D., Jean-Pierre Marolleau M.D., Ph.D., Marc Polivka M.D., Clovis Adam M.D., Renata Ursu M.D., Anna Schmitt M.D., Noemie Barillot Ms.C., Lucia Nichelli M.D., Ms.C., Fernando Lozano-Sánchez, M.D., Maria-José Ibañez-Juliá M.D., Matthieu Peyre M.D., Ph.D., Bertrand Mathon M.D., Ms.C., Yah-se Abada Ph.D., Frédéric Charlotte M.D., Ph.D., Frédéric Davi M.D., Ph.D., Chip Stewart Ph.D., Aurélien de Reyniès Ph.D., Sylvain Choquet M.D., Carole Soussain M.D., Ph.D., Caroline Houillier M.D., Bjoern Chapuy M.D., Ph.D., Khê Hoang-Xuan M.D. Ph.D., Agustí Alentorn, M.D., Ph.D.*

[1]Institut du Cerveau-Paris Brain Institute-ICM, Inserm, Sorbonne Université, CNRS, F-75013 Paris, France

[2]Department of Hematology, Oncology and Cancer Immunology, Campus Benjamin Franklin, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, 12203, Berlin, Germany

[3]Department of Neuropathology, Groupe Hospitalier Pitié Salpêtrière, APHP, F-75013, Paris, France

[4]Department of Pathology, CHU de Bordeaux, Hôpital Pellegrin, 33076, Bordeaux, France titut du Cerveau-Paris Brain Institute-ICM, Inserm, Sorbonne Université, CNRS, APHP, Hôpital de la Pitié Salpêtrière, DMU Neurosciences, Service de Neurologie 2-Mazarin, F-75013 Paris (Y.S.A., F.L-S., C.H., K.H.X. and A.A.), Department of Neuroradiology, Sorbonne Université, Assistance Publique-Hôpitaux de Paris, Groupe Hospitalier Pitié-Salpêtrière-Charles Foix (L.N.), Sorbonne Université, Paris Brain Institute (ICM; INSERM, UMRS 1127; CNRS, UMR 7225), Paris and AP-HP, Service de neurochirurgie, Hôpital de la Pitié-Salpêtrière, Paris (M.P and B.M.), Department of Pathology, APHP, Hôpital Pitié-Salpêtrière and Sorbonne University, Paris (F.C.), Département de Neuropathologie, Paris (K.M.), from Department of Hematology, APHP, Hôpital Pitié-Salpêtrière and Sorbonne University, Paris (F.D. and S.C.) from Department of Neurosurgery, Bordeaux University Hospital Center, Pellegrin Hospital, Place Amélie Raba-Léon, 33076, Bordeaux and EA 7435 - IMOTION, University of Bordeaux, Bordeaux (H.L.), Service de Pathologie, CHU de Bordeaux, Hôpi-

tal Pellegrin, 33076, Bordeaux (S.E), Département de pathologie, PBH, CHU Angers, 4, rue Larrey, 49933 Angers cedex 9, France; CRCINA, université de Nantes - université d'Angers, Angers (A.R.), Department of Hematology, CHU Angers, 4, rue Larrey, 49933 Angers cedex 9, France (J.P.); Department of Neurology, Hôpitaux Civils de Colmar, Colmar, France (G.A.), Department of Neuropathology, Hôpitaux civils de Colmar, 68000 Strasbourg, France (F.L.); Department of Pathology, CHU de Toulouse, IUC-Oncopole, 31300 Toulouse, INSERM U1037, Cancer Research Center of Toulouse (CRCT), 31100 Toulouse, Université Toulouse III Paul Sabatier, 31062 Toulouse (E.U.C), Department of hematology, IUC Toulouse Oncopole, Toulouse (L.O.); Neuropathology Department University Hospital Timone, Aix Marseille University, Marseille and Inst Neurophysiopathol, CNRS, INP, Aix-Marseille University, Marseille (D.F.B.), APHM, CHU Timone, Service d'anatomopathologie, Marseille, France (O.C.), Department of Biopathology, CHRU Nancy, CHRU/ICL, Bâtiment BBB, Rue du Morvan, 54511, Vandoeuvre-lès-Nancy and Department of Legal Medicine, CHRU Nancy, Vandoeuvre-lès-Nancy and INSERM U1256, University of Lorraine, Vandoeuvre-lès-Nancy and Centre de Ressources Biologiques, BB-0033-00035, CHRU, Nancy (G.G.), Department of Neuro-Oncology, CHRU-Nancy, Université de Lorraine, Nancy, France (L.T.), Department of Hematology, Centre Hospitalier Universitaire Amiens-Picardie, Amiens (J.-P. M.); Department of Anatomopathology, Laribosière Hospital, Assistance Publique - Hopitaux de Paris, University of Paris, Paris, France (M.P.); Pathology Department, Bicêtre University Hospital, Public Hospital Network of Paris, Le Kremlin Bicêtre, France (C.A.); Department of Neurology, Université de Paris, AP-HP, Hôpital Saint Louis, 75010, Paris, France (R.U.); Department of Hematology, Institut Bergonié Hospital, Bordeaux, France (A.S.); Neurology Department, Perpignan hospital, Perpignan, from INSERM UMR_S1138 - Centre de Recherche des Cordeliers - Université Pierre et Marie Curie et Université Paris Descartes, Paris (A.D.R.), Center for Cancer Immunotherapy, Institut Curie, PSL Research University, INSERM U932, 75005 Paris and Clinical Hematology Unit, Institut Curie, 92210 Saint-Cloud (C.S.), all in France. From Department of Hematology, Oncology and Cancer Immunology, Campus Benjamin Franklin, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, 12203, Berlin (K.W. and B.C.) all in Germany. From Broad Institute of MIT and Harvard, Cambridge, MA 02142 (C.S.), in the USA.

*Corresponding author:

Agusti Alentorn, MD PhD

Email address: agusti.alentorn@icm-institute.org

Mailing address: Department of Neurology-2, Mazarin, Groupe Hospitalier Pitié Salpêtrière, 75013 Paris, France

Phone: +33 142164160

Abstract

Primary central nervous system lymphoma (PCNSL) is a rare and distinct entity within diffuse large B cell lymphoma with variable response to treatment and outcome despite homogenous pathological presentation. The clinical impact of its molecular aberrations and the presence of molecular heterogeneity are poorly understood.

We performed genome-wide analysis of 147 PCNSL from fresh-frozen tumor tissue from immunocompetent, treatment naïve patients, incorporating exome sequencing, DNA copy number, DNA methylation, and RNA expression. These data were integrated and correlated with the clinico-radiological characteristics and outcomes of the patients. We validated our results in an independent series of 93 PCNSL formalin-fixed, paraffin-embedded (FFPE) samples.

Consensus clustering of multi-omics data uncovered concordant classification of four robust, non-overlapping, prognostically significant clusters (CS). The CS1 group, characterized by high proliferation and PRC2 complex activity, had an intermediate outcome between CS2/CS3 and CS4. Patients who had PCNSL with an "immune-hot" (CS4) profile had the most favorable clinical outcome. However, the immune-cold hypermethylated CS2 and the heterogenous-immune CS3 groups had a poor prognosis. Nearly all PCNSL patients with meningeal infiltration harbored *HIST1H1E* mutations, enriched in the CS3 group. The integrated analysis suggests that the CS4 group may be more susceptible to immunotherapy than the other groups. The integration of genome-wide data from multi-omics data revealed four molecular patterns in PCNSL with a distinctive prognostic impact that significantly improved the current clinical stratification. This molecular classification using FFPE samples facilitates routine use in clinical practice and provides potential precision-medicine strategies in PCNSL.

5.3.1 Introduction

Primary central nervous system lymphoma (PCNSL) is a rare subtype of extranodal non-Hodgkin's lymphoma within diffuse large B cell lymphoma (DLBCL), but has a less favorable prognosis that its systemic counterpart and has been proved to be molecularly a different biological entity (Chapuy et al., 2018; Schmitz et al., 2018; Sehn & Salles, 2021; Yoshida et al., 2016). The standard treatment relies on high-dose methotrexate (HD-MTX) regimen with or without consolidation and is associated with treatment resistance or relapses in up to 60% of the patients (Houillier et al., 2020; Y. Zhou et al., 2018).

Biologically, initial studies have found PCNSL to be at late B-cell germinal center (GC) exit stages and to have constitutive NF-$\kappa\beta$ activity driven by mutations in genes of the B-cell receptor (BCR) pathway, of the toll-like receptor (TLR) pathway (*MYD88*) and *CARD11*. Recently, DLBCL has been divided in different molecular clusters and PCNSL have been related to the so-called "MCD" (based on the co-occurrence of $MYD88^{L265P}$ and *CD79B* mutations) or Cluster 5 (C5) DLBCL, both converging in the presence of frequent $MYD88^{L265P}$, *CD79B*, *PIM1*, *BTG2* mutations, IgH-BCL6 translocations, copy gains of 3q12.3, 9p24.1 (PD-L1/PD-L2), and copy losses of 6p21-22 (HLA locus), 6q21, and 9p21.3 (*CDKN2A* biallelic loss) (Chapuy et al., 2016, 2018; Fukumura et al., 2016; Schmitz et al., 2018). Currently, PCNSL heterogeneity has not been properly addressed mainly due to the lack of multi-omic data integration and the limited number of patients (Fukumura et al., 2016).

Here, we performed an integrative analysis of mutations, copy-number alterations (CNA), fusions, gene expression, TCR/BCR clonotypes, tumor microenvironment (TME), methylation, tumor localization, and clinical data to identify molecular subtypes of PCNSL with clinically distinct behaviors. Additionally, to facilitate routine clinical implementation, we developed an algorithm that uses gene expression data from either formalin-fixed, paraffin-embedded (FFPE), or fresh-frozen (FF) tissue, to identify the PCNSL molecular subtypes associated with multi-omic features.

5.3.2 Methods

Patients

A total of 147 FF (discovery cohort) and 93 FFPE (validation cohort) tumor samples from immunocompetent Epstein-Barr negative PCNSL were recollected from different French hospitals (see Table S1 in Supplementary Appendix 1) after written informed consent and ethics approval (Pitié Salpêtrière Hospital ethics committee) were obtained. All the tumors were newly diagnosed PCNSL, pathologically confirmed according to the World Health Organization classification and

were treated with standard HD-MTX based chemotherapy regimen according to LOC network recommendations (Houillier et al., 2020).

Sequencing platforms and multi-omic data integration for PCNSL molecular subtyping

We performed, on the FF cohort, exome sequencing (n=115) to call mutations and obtain CNA events, RNA sequencing (n=123) to analyze gene expression, immune cell proportions, TCR/BCR clonotypes, and fusion transcripts, and DNA methylation profiling (n=64). High robust clustering was obtained by consensus clustering resulting from 10 different multi-omics clustering algorithms that are integrated in the R package "MOVICS" (Lu, Meng, Zhou, Jiang, & Yan, 2021).

Statistical analyses

Differences in proportions and binary/categorical variables were calculated from Fisher's exact test. Kruskal-Wallis test was used to test for a difference in distribution between three or more independent groups, and Mann Whitney U test was used for differences in distributions between two population groups unless otherwise noted. Overall Survival (OS) analysis was assessed using log-rank Kaplan-Meier curves and multivariate Cox proportional hazards regression modeling. See Supplementary Appendix 1 for full details.

5.3.3 Results

Multi-omic data integration reveals PCNSL molecular subtypes with clinical outcome implications

We performed a cluster of clusters analysis using six levels of omic information (Figure 5.9A and Fig. S1-S3) to identify four PCNSL subtypes (CS1 to CS4) that display different clinical outcomes in OS (Global log-rank $p < 0.001$, Figure 5.9B). Patients in CS4 had the longest OS ($median = 66.8$ months; 95% confidence interval [CI]$= 19.8 - 67.2$) and lived significantly longer than those in both clusters CS2 (median=18 months; $CI_{95\%} = 8.3 - 53.4$; $p = 0.024$) or CS3 ($median = 13.8$ months; $CI_{95\%} = 6.1 - 16.7$; $p = 0.003$), and slightly longer, but not significantly, to those in CS1 ($median = 26.2$ months; $CI_{95\%} = 13.3 - 63.9$; $p = 0.094$). Additionally, these observations remained significant after adjusting by age and Karnofsky Performance Status (KPS) in Cox proportional hazard ratio multivariate models (Fig. S4A). Interestingly, CS4 was independently associated with a better response when considering progression free survival in univariate and multivariate models (Fig. S4B-C). Finally, we did not observe significant differences in the median number of predicted immunogenic neoantigens ($p = 0.44$, Table S2 and Fig. S5).

Transcriptomic data correctly assign multi-omic defined PCNSL subtypes in FF and FFPE samples

Given the difficulty of acquiring FF tissue and of analyzing and implementing multi-omic data into routine clinical practice, we sought to evaluate the use of only RNA expression, obtained from FFPE or FF tissue, to categorize patients into the four PCNSL CS (Table S3). We obtained a Cohen's kappa coefficient of 0.90 ($p < 0.001$) when evaluating the accuracy of correctly assigning patients from the multi-omic cohort. Additionally, when expanding to the FF-RNA complete set or when using the FFPE cohort, we observed the same behaviors regarding clinical outcome (Global $p < 0.001$) across molecular subtypes in both univariate and multivariate models (Figure 5.9C and Fig. S6-S12).

Next, we evaluated the contribution of each omic-level data to outcome prediction models by using Harrell's concordance index (C-index) (Goeman, 2009). A C-index of 0.60 ($0.56 - 0.65$ at $CI_{95\%}$) in FF and 0.71 ($0.68 - 0.74$ at $CI_{95\%}$) in FFPE was observed using KPS and age, which are the clinical features currently used in the Memorial Sloan Kettering Cancer Center prognostic score for PCNSL (Abrey et al., 2006). When adding different omic-data to the FF cohort modeling, we observed higher predictive power using mRNA expression compared to the other omic data (C-index $= 0.91 \pm 0.02$ at $CI_{95\%}$). We further validated these observations in the FFPE cohort obtaining a C-index of 0.83 ($0.80 - 0.85$ at $CI_{95\%}$) and 0.93 ($0.91 - 0.95$ at $CI_{95\%}$) when adding the mRNA level or the TME and RNA levels to the model, respectively (Figure 5.9D). Altogether, these results show that RNA-seq data from FFPE or FF tissue can be used to correctly identify PCNSL subgroups.

Mutational landscape of PCNSL

We identified $32,544$ mutations in the 115 PCNSL samples analyzed ($median =$ 3.23 mutations/Mb; $range = 0.02 - 85.49$; Table S4 and Fig. S13). We applied the dNdScv (Martincorena et al., 2017) algorithm to identify driver mutations identifying the hallmark mutations of PCNSL like *MYD88* (64%), *PIM1* (59%), *PRDM1* (57%), *GRHPR* (50%), *HLA-A/B/C* (49%, 30%, and 13%), *BTG2* (47%), *CD79B* (43%), *CDKN2A* (28%), *TBL1XR1* (25%), *KLHL14* (25%), *CARD11* (22%), and *HIST1H1E* (18%) which are involved in BCR-TLR mediated NF-$\kappa\beta$ signaling, antigen presentation, cell-cycle, histone modification and B-cell differentiation regulation (Bruno et al., 2014; Chapuy et al., 2016; Fukumura et al., 2016) (Figure 5.10A, Table S5 and Fig. S14). Moreover, we detected canonical activation-induced cytidine deaminase (c-AID) off-target mutations and found they represent 7.9% ($6.8 - 8.5\%$ at 95% CI) of SNV mutations and fall within driver genes like *PIM1* (47%), *CD79B* (10%), *IRF4* (9%), and *HIST1H1E* (6%) (Table S6,

Fig. S15-S16). Interestingly, both c-AID and non c-AID (Cosmic signature SBS9) mutations are significantly more active at clonal stages ($p = 0.007$ and 0.018, respectively), hence reflecting the importance of AID activity in the early stages of PCNSL tumorigenesis (Fig. S17-18).

Regarding focal CNA, we identified significant recurrent amplifications in 18q21.33 (42%), and 19p13.13 (34%), and deletions in 6p21 (39%), 6q21 (65%), 6q27 (49%), and 9p21.3 (28%) which have a higher frequency than those observed in systemic DLBCL (Figure 5.10B) (Chapuy et al., 2016, 2018). Furthermore, we found additional, not previously described, amplifications in 1q32.1 (33%, *IL10*), and 11q23.3 (26%, *CD3G*), and deletions in 6p25.3 (21%, *IRF4*), 22q11.22 (29%, *GGTLC2*) and 14q32.33 (84%) that produce significant expression changes in *CD3G* ($FC = 1.25$), *IRF4* ($FC = -1.03$) and *GGTLC2* ($FC = -1.76$, FDR q-value < 0.1), respectively (Table S7, Fig. S19-S22).

Distinct genetic signatures within PCNSL subtypes and systemic DLBCL

Afterwards, we aimed to characterize the differences in genetic alterations across groups for each mutation, focal CNA, and fusion. The CS4 cluster presents ten enriched events that included mutations in *SOCS1*, which is a negative regulator of the JAK-STAT3 pathway, *MPEG1*, *PIM2*, and deletion of 17q25.1 involving *GRB2* that indirectly regulates the NF-$\kappa\beta$ pathway. We observed 43 events within the CS1 cluster including mutations involved in NF-$\kappa\beta$ pathway (*RIPK1* via 6p25.3 deletion), B-cell differentiation (*IRF4* via 6p25.3 deletion, *TOX*, and *BCL6*), proliferation via interruption of cell cycle arrest (*CDKN2A/2B* fusions and *FOXC1*), and B-cell lymphomagenesis (e.g., *ETV6*, *OSBPL10*). Patients within the CS3 cluster exhibit 12 events from which *HIST1H1E* arises as the top enriched, and has been proved to enhance self-renewal properties and disrupt chromatin architecture in B-cell lymphomas (Chapuy et al., 2016, 2018; Schmitz et al., 2018; G. W. Wright et al., 2020; Yusufova et al., 2021). The CS2 cluster did not present any genomic characteristic events. Furthermore, most of these distinctive events arrived as early events (clonal) in tumorigenesis like *IRF4* and *BCL6* in CS1 (Figure 5.10C, Table S8-S9). Of note, most of these mutations were not observed in the clusters previously defined by Chapuy et al. (e.g., 9p11.2 del; Figure 2D) (Chapuy et al., 2018).

B-cell differentiation stages, pathways, and TME distinctions between PCNSL molecular subtypes

We recovered and analyzed the expression of different previously curated gene signatures (Schubert et al., 2018; G. W. Wright et al., 2020) (see Methods). CS1 was

characterized by the upregulation of PI3K, glycolytic activity, and cell proliferation signatures; additionally, it presented hyperactivation of the PRC2 complex which has been proved to inhibit MHC-I expression, through histone methylation (Figure 5.11A, $p < 0.05$, and Fig. S23-S25) (Fangazio et al., 2021). Moreover, p53 activity was enriched in the CS2 cluster (Holland, Szalai, & Saez-Rodriguez, 2020; Schubert et al., 2018). Interestingly, even though all clusters presented mutations within the NF-$\kappa\beta$ pathway, it was transcriptionally active only in clusters CS3 and CS4. Additionally, MAPK and JAK-STAT pathways were upregulated in those clusters, respectively (Figure 5.11A).

Regarding B-cell differentiation programs, CS1 expressed a mixture of GC cells which is consistent with the 6p25.3-19q13.12 deletions, and *BCL6* mutations (Figure 5.10C). On the other hand, cluster CS4 presents an enrichment in terminally differentiated plasma cells that goes in line with *BCL6* downregulation, the absence of *MYC* induction, and *BCL6* mutations. The most heterogeneous cluster was the CS3, presenting features of both GC and mature B-cells (plasma cells and memory B-cells). Intriguingly, the cluster CS2 did not present any B-cell stage enrichment but instead a lymphatic endothelial cell (LEC) gene signature (Figure 5.11A and Fig. S26-27).

Then, we aimed to describe the TME differences between subtypes by using CIBERSORTx derived immune deconvolution and B-cell lymphoma specific TME gene signatures (Kotlov et al., 2021). CS1 cluster is immunologically "neutral" meanwhile the CS2, which is immunologically depleted, exhibits expression of vascular endothelial cells (VEC), memory resting CD4+ T-cells, monocytes, and activation of GABA synthesis, which has been recently linked to B-cells that inhibit CD8+ T-cells' killer function and promote monocyte differentiation into anti-inflammatory macrophages (Baihao Zhang et al., 2021). The CS4 cluster has a hot-inflammatory TME due to the presence of active CD8+ T-cells and NK cells (with high cytolytic activity score) (Rooney, Shukla, Wu, Getz, & Hacohen, 2015). Conversely, heterogeneity was again observed for the CS3 subtype, being only inactivated macrophages M0 more significantly enriched (Figure 5.11A, Fig. S28-S41, and Table S9-S10).

CS3 subtype is associated with meningeal infiltration to cerebrospinal fluid

Here, we investigated if brain MRI analysis ($n = 90$, FFPE cohort) could provide more insights on the molecular subtypes. We observed no brain lobe preference between PCNSL subgroups but, in general, tumors arose less in the occipital lobe (4/90 cases versus 86/90, $p < 0.001$). In addition, CS4 tumors arose more in the isthmus of the corpus callosum (7/34 cases versus 0/56, $p < 0.001$). Conversely, CS2/CS3 were more frequent in the brainstem (4/16 and 3/19 cases versus 1/55,

$p = 0.005$), when compared to the other clusters. Strikingly, we found no association with tumor size nor multiple lesions. However, meningeal infiltration of the cerebrospinal fluid (CSF) was only found within CS3 tumors (6/16 cases versus 0/74, $p < 0.001$, Figure 5.11B and Table S11).

Epigenetic attributes across PCNSL subtypes

We proceeded to investigate epigenetic differences among subtypes ($n = 64$). The CS2 displayed higher hypermethylation globally ($p = 0.006$, Fig. S42-S43). Interestingly, GO analyses on differentially methylated promoters revealed B-cell differentiation programs to be hypomethylated in CS1 but hypermethylated in CS2; while interleukin-1 was hypermethylated in CS4 (Figure 5.12A and Table S12). Genomic region enrichment analysis on hypermethylated promoters identified strong enrichment of binding sites for the histone/chromatin proteins H3K27me3 and EZH2 in CS1, and NF-$\kappa\beta$, *IRF4*, and *BCL6* in CS2 (Figure 5.12B, Fig. S44 and Table S13).

From multi-omics to potential therapeutic targets

To generate an explanatory bridge between the different multi-omic layers and ultimately potential therapeutic targets across subtypes, we integrated all multi-omic data and evaluated their contribution to specific pathways. Even though the hallmark PCNSL alterations targeting My-T-BCR protein supercomplex, CD79A/B BCR subunits, TNFAIP3, RIPK1, TAB2, and the CBM (CARD11-BCL10-MALT1) complex were relatively constant across subgroups, the NF-$\kappa\beta$ hyperactive group (CS4) presented more *GRB2/LYN* deletions and absence of *PLCG2* mutations, which either represses the BCR complex or affects the CBM complex activation. Furthermore, NF-$\kappa\beta$ activity could not be explained by self-antigen-dependent chronic active BCR signaling upregulation since $\text{IgV}_{\text{H4-34}}$ expression was similar across groups (Figure 5.13) (Jang et al., 2011; Phelan et al., 2018; G. W. Wright et al., 2020). These observations suggest that CS4 and CS3 may be more sensitive to BTK inhibitors (e.g., ibrutinib). The CS4 cluster also presented high JAK-STAT activity and mutated *SOCS1* (a JAK1 repressor), making it potentially responsive to JAK1 inhibitors (e.g., INCB040093) (Linossi & Nicholson, 2015; Phillips et al., 2018). Regarding antigen presentation-related genes, we observed only monoallelic deletions in *HLA-A*, *B2M*, and *CD58* but not in *HLA-B* or *HLA-C*. Moreover, the absence of PRC2 complex activity and presence of MHC-I and checkpoint molecules expression indicate a potential use of immune checkpoint inhibitors (ICI) for CS4. On the other hand, EZH2 inhibitors (e.g., tazemetostat) in combination with ICI could potentially increase MHC-I expression and immune detection in CS3 (Dersh et al., 2021). Interestingly, the

CS3 cluster is enriched with *HIST1H1E/C* mutations which have been recently demonstrated to confer enhanced fitness, and self-renewal properties to B-cells (Yusufova et al., 2021).

Additionally, we observed a higher frequency of cases with genetic alterations involved in the cell cycle for CS1 (97%, $p < 0.001$, e.g., *CDKN2A/2B* fusions); hence, cyclin D-Cdk4,6 plus PI3K inhibitors could be beneficial for CS1 patients.

Despite not presenting enriched genetic signatures, the CS2 cluster may be potentially susceptible to inhibition of the TFs IRF4 (e.g., lenalidomide), SPIB, and MEIS1 (e.g., MEISi-1), and/or inhibition of GAD67 (G. W. Wright et al., 2020; Baihao Zhang et al., 2021).

5.3.4 Discussion

Identifying groups of patients with shared biologic and prognostic markers is extremely challenging mainly due to high genetic, phenotypic and TME heterogeneity. We identified four PCNSL molecular subtypes with specific oncogenic pathways, gene expression phenotypes, methylation profiles, TME, tumor location, outcome, and potential therapeutic targets (Figure 5.13). Moreover, our study gives plausible explanations to the PCNSL response heterogeneity based on finding that many previously PCNSL characteristic features, based on MCD or C5 DLBCL subtypes (Chapuy et al., 2018; G. W. Wright et al., 2020), are cluster-specific (C1-C4) and not shared across all PCNSL tumors. For example, PCNSLs (MCD/C5 DLBCLs) are mainly characterized by mutations leading to constitutive NF-$\kappa\beta$ activation but this was only observed, transcriptionally, for CS4 and CS3; however, the outcome for these clusters is very different mainly due to tumor location, TME, and B-cell differentiation program differences.

Moreover, we propose different tailored treatments according to the pathway activation of each CS, suggesting, for example, that CS4 might be more likely to respond to ICI treatment.

On top of this and given the importance of routine clinical implementation, we propose RNA-based Brain Lymphoma Profiler (RBraLymP), which uses gene expression data from either FFPE or FF tissue, to identify the PCNSL molecular subtypes associated with multi-omic features. The RBraLymP algorithm is publicly accessible at `https://github.com/iS4i4S/PCNSL-RBraLymP` such that existing and new therapy efforts can be directed to the most appropriate patients.

In summary, our multi-omics analysis builds on the current classification of DLBCL by the addition of the molecular heterogeneity within PCNSL that may inform on its pathogenesis. Our study discovered a link between genetic and neoplastic signaling pathways, pointing to potential treatment targets. Selecting treatment for PCNSL based on individual genetic changes is not desirable from the standpoint of precision medicine, as it is likely that combinations of genetic

aberrations influence therapeutic response. The genetic subgroups we define could serve as a conceptual foundation for developing targeted therapeutic approaches for these poorly understood and with high mortality malignancies.

5.3.5 Acknowledgments

We thank Yannick Marie and Emeline Mundwiller (Paris Brain Institute), Lucile Armenoult and Mira Ayadi (Carte d'Identité de Tumeurs, Ligue contre le Cancer) for the help with exome, RNA sequencing and bisulfite DNA sequencing libraries preparation; Silvia Duran for helpful discussions; Figure 5.13 was created with biorender.com and all the clinicians, pathologists, researchers and patients associated with the French National "Lymphome Oculo-Cérébral" included in Rare cancers of the central nervous system, RENOCLIP-LOC Network, approved by the French National Institute of Cancer (INCa).

5.3.6 Funding

This work was in part supported by a grant from Investissements d'avenir and by the grant INCa-DGOS-Inserm_12560 of the SiRIC CURAMUS, the program "investissements d'avenir" ANR-10-IAIHU-06, PRT-K/INCa grant LOC-model reference 2017-1-RT-04, DRCI de l'APHP, CRC2013_105_R1 / projet Tri LOC, BETPSY project, overseen by the French National Research Agency, as part of the second "Investissements d'Avenir' ' program (Grant No. ANR-18-RHUS-0012), Foundation RAM active investments, ARTC foundation, an unrestricted grant from Bristol Myers Squibb (BMS): RDON06618, ICGex project and IDeATIon project with an unrestricted grant from MSD Avenir.

5.3.7 Declaration of interest

G.A reports grants from Biogen, Novartis, Roche, Sanofi, Abbvie, Pfizer and CSL Behring, outside the submitted work. A.A. reports research grant with an unrestricted grant from Bristol Myers Squibb (BMS).

5.3.8 Supplementary Material

Full methods description (including statements of data availability and any associated accession codes and references) and supplementary figures are available in the Appendix 1. Most supplementary tables will only be available in the online version of the paper.

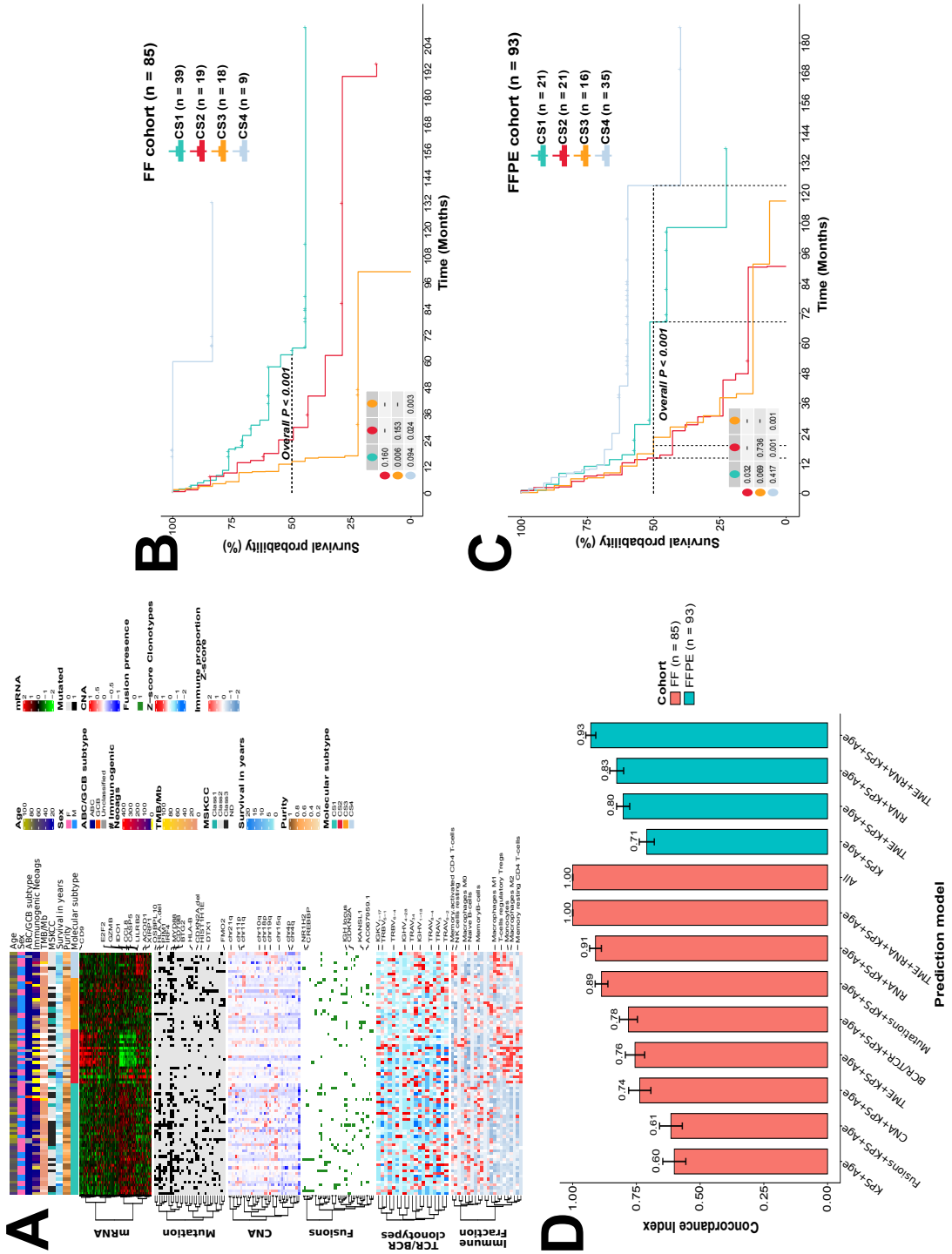

A
B
C
D
FF cohort (n = 85)
CS1 (n = 39)
CS2 (n = 19)
CS3 (n = 18)
CS4 (n = 9)
Overall P < 0.001
Survival probability (%)
Time (Months)
FFPE cohort (n = 93)
CS1 (n = 21)
CS2 (n = 21)
CS3 (n = 16)
CS4 (n = 35)
Overall P < 0.001
Survival probability (%)
Time (Months)
Cohort
FF (n = 85)
FFPE (n = 93)
Concordance Index
Prediction model
KPS+Age
Fusions+KPS+Age
CNA+KPS+Age
TME+KPS+Age
BCR/TCR+KPS+Age
Mutations+KPS+Age
RNA+KPS+Age
TME+RNA+KPS+Age
All
KPS+Age
TME+KPS+Age
RNA+KPS+Age
TME+RNA+KPS+Age
0.60
0.61
0.74
0.76
0.78
0.89
0.91
1.00
1.00
0.71
0.80
0.83
0.93
mRNA
Mutation
CNA
Fusions
TCR/BCR clonotypes
Immune Fraction

Figure 5.9 *(preceding page)*: Multi-omic data integration reveals PCNSL molecular subtypes with clinical outcome implications. Panel A shows the resulting consensus heatmap based on 10 integrative clustering algorithms to define the clusters (CS1 to CS4) where each of the 10 algorithms uses cluster of clusters analysis to integrate six levels of omic information (y-axis) in the order: i) mRNA expression (2,087 variables), ii) mutations (31 variables), iii) CNA (40 variables), iv) fusion transcripts (43 variables), v) TCR/BCR clonotypes (19 variables), and vi) immune cell fractions (22 variables). Additional genomic and clinical features are annotated at the top. Panel B shows Kaplan-Meier estimates of overall survival among patients belonging to each cluster that resulted from the Consensus cluster of clusters analysis. Panel C shows Kaplan-Meier estimates of overall survival among patients belonging to each cluster using an FFPE validation cohort (n=93). Age and KPS multivariate models for both cohorts are shown in Fig. S4-S13 in Supplementary Appendix 1. Panel D shows the Harrell's concordance index (value annotated at top of each bar) obtained when evaluating each omic-level data to outcome prediction models using Cox proportional hazards regression. The prediction was overfit when using ALL omic data on the FF cohort. Observations were validated (same direction and effect) using RNA and TME data from the FFPE cohort. Error bars indicate the 95% confidence intervals.

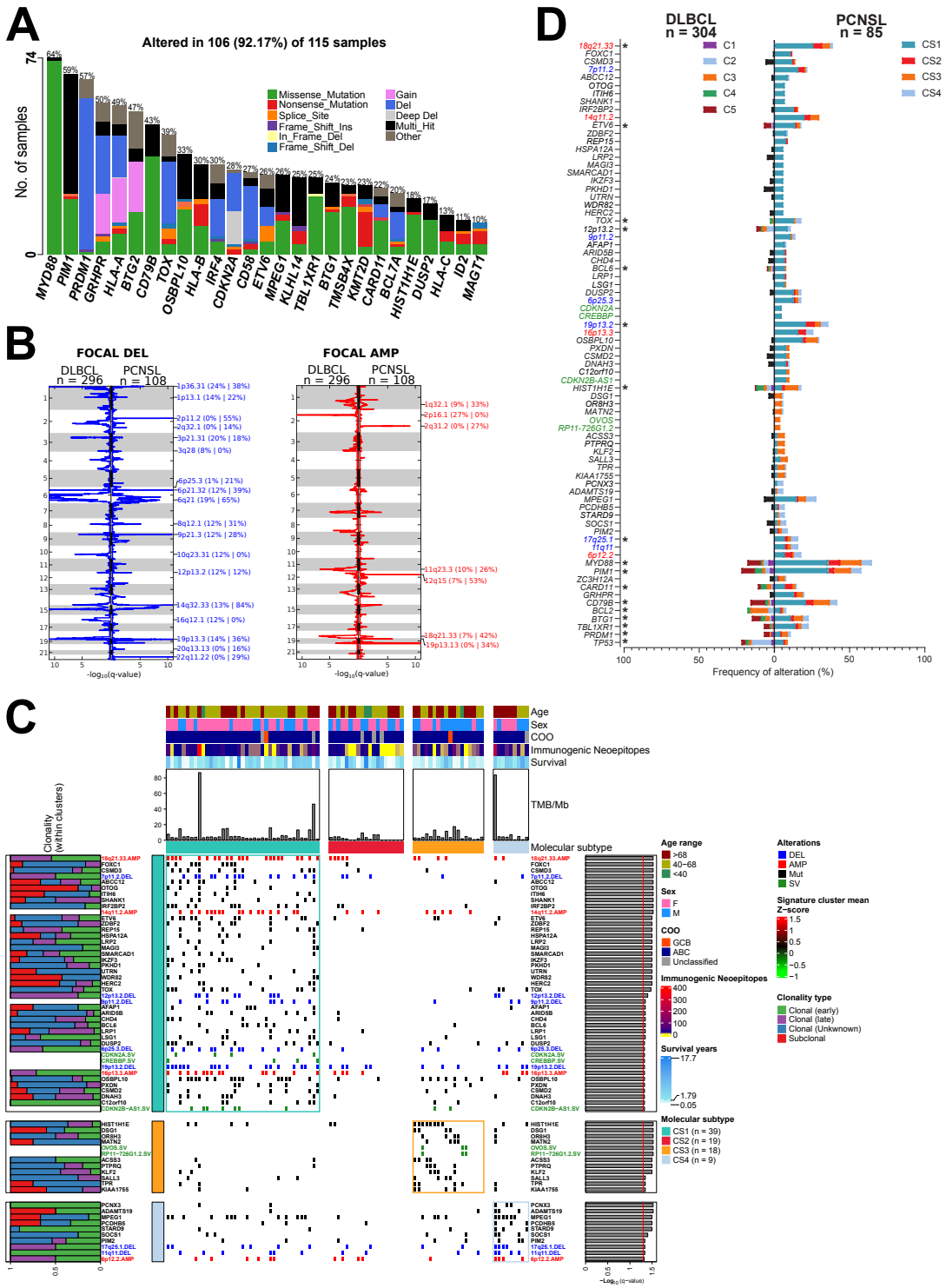

A
Altered in 106 (92.17%) of 115 samples
No. of samples
Missense_Mutation
Nonsense_Mutation
Splice_Site
Frame_Shift_Ins
In_Frame_Del
Frame_Shift_Del
Gain
Del
Deep Del
Multi_Hit
Other
MYD88 PIM1 PRDM1 GRHPR HLA-A BTG2 CD79B TOX OSBPL10 HLA-B IRF4 CDKN2A CD58 ETV6 MPEG1 KLHL14 TBL1XR1 BTG1 TMSB4X KMT2D CARD11 BCL7A HIST1H1E DUSP2 HLA-C ID2 MAGT1
B
FOCAL DEL
DLBCL n = 296
PCNSL n = 108
FOCAL AMP
-log10(q-value)
C
Age
Sex
COO
Immunogenic Neoepitopes
Survival
TMB/Mb
Molecular subtype
Clonality (within clusters)
Age range
Sex
COO
Immunogenic Neoepitopes
Survival years
Molecular subtype
Alterations
Signature cluster mean Z-score
Clonality type
D
DLBCL n = 304
PCNSL n = 85
C1
C2
C3
C4
C5
CS1
CS2
CS3
CS4
Frequency of alteration (%)

Figure 5.10 *(preceding page)*: Distinct genetic signatures within PCNSL subtypes and systemic DLBCL. Panel A shows the number of affected samples within the top driver genes (identified by dNdScv algorithm) in the cohort of 115 PCNSL patients. Barplots are filled according to mutation type (missense, nonsense, splice site, frameshift, multihit, or other) or CNA events (gain, deletion, or deep deletion). The frequency of affected samples within the cohort is annotated at the top of each barplot for each driver gene. Panel B shows the GISTIC2.0-defined recurrent copy number focal deletions (blue, left) and gains (red, right) as mirror plots in DLBCL (n=296 from Chapuy et al., 2018) and PCNSL (n=108 from this study). Chromosome position is on the y-axis, and significance is on the x-axis. CNAs are labeled with their associated cytoband/arm followed in brackets by the frequency of the alteration (DLBCL | PCNSL). Panel C shows the landmark genetic alterations for each PCNSL subtype (boxed for each cluster) identified by a one-sided Fisher test (event within-cluster vs outside-cluster) and ranked by significance (FDR corrected q-value ≤ 0.1 selected, red line, bar plot to the right). The left bar plot shows the relative contribution of temporal acquisition for each alteration event (only within the enriched cluster) to indicate how early or late during tumorigenesis the event might have happened. Additional genomic and clinical features are annotated at the top. COO, cell of origin; F, female; M, male; DEL, deletion; AMP, amplification; Mut, mutation; Fusion, fusion transcript. Panel D shows a mirror bar plot with the frequencies of recurrent genetic alterations in PCNSL's clusters (n=85) compared to those in DLBCL's clusters (n=304, Chapuy et al., 2018). Asterisks denote the known driver events in DLBCL and colors the alteration type (mutation = black; gain = red; loss = blue; structural variant = green).

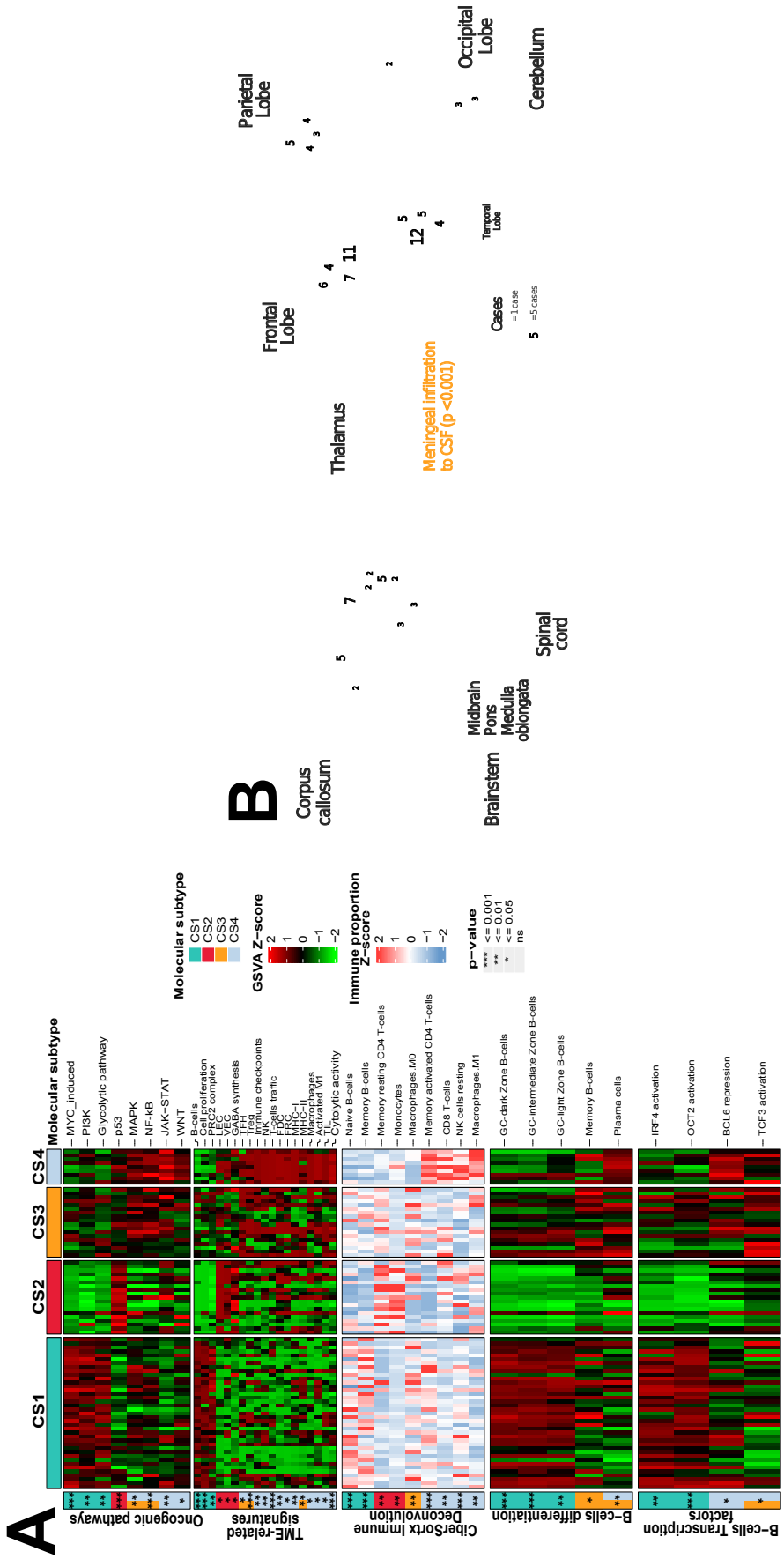
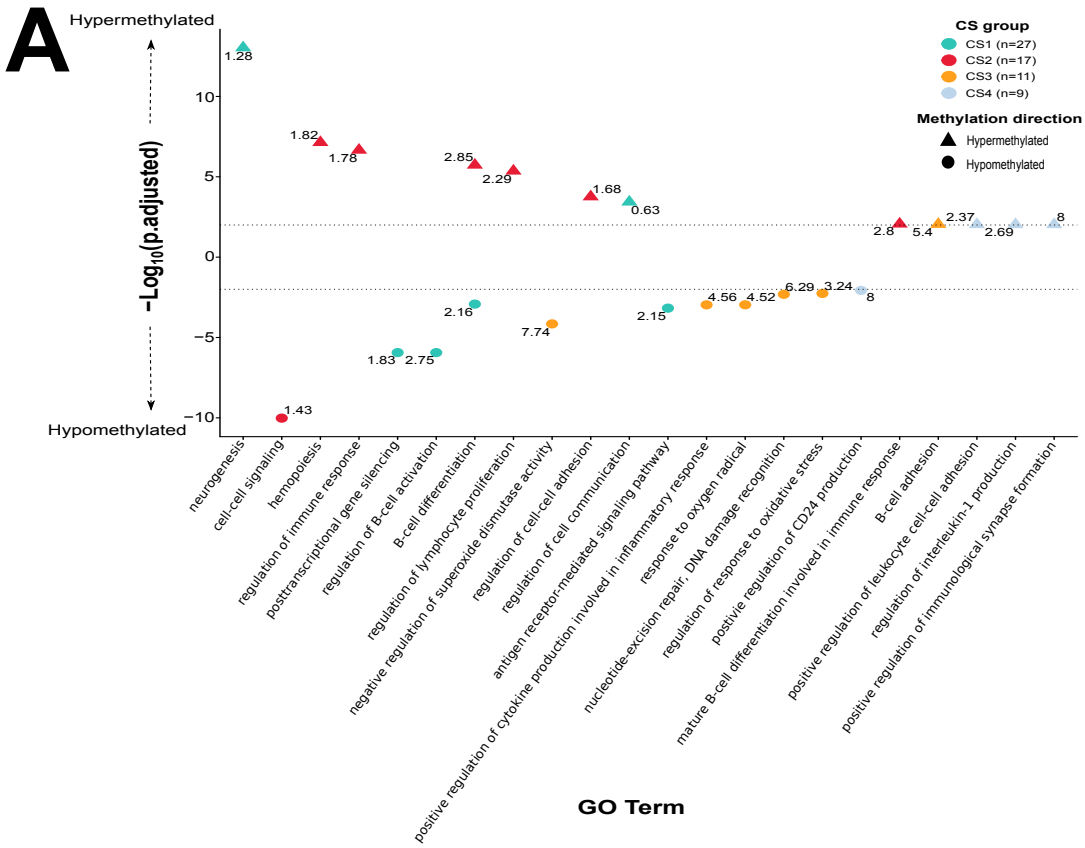

A
CS1
CS2
CS3
CS4
Molecular subtype
MYC_induced
PI3K
Glycolytic pathway
p53
MAPK
NF-kB
JAK-STAT
WNT
Oncogenic pathways
B-cells
cell proliferation
PRC2 complex
VEC
CAF
GABA synthesis
Treg
Immune checkpoints
T-cells traffic
FRC
MHC-I
MHC-II
Macrophages
Activated M1
Cytolytic activity
TME-related signatures
Molecular subtype
CS1
CS2
CS3
CS4
GSVA Z-score
2
1
0
−1
−2
Naive B-cells
Memory B-cells
Memory resting CD4 T-cells
Monocytes
Macrophages M0
Memory activated CD4 T-cells
CD8 T-cells
NK cells resting
Macrophages M1
CIBERSORTx Immune Deconvolution
Immune proportion Z-score
2
1
0
−1
−2
p−value
*** <= 0.001
** <= 0.01
* <= 0.05
ns
GC-dark Zone B-cells
GC-intermediate Zone B-cells
GC-light Zone B-cells
Memory B-cells
Plasma cells
B−cells differentiation
IRF4 activation
OCT2 activation
BCL6 repression
TCF3 activation
B−cells Transcription factors
B
Corpus callosum
Thalamus
Frontal Lobe
Parietal Lobe
Occipital Lobe
Temporal Lobe
Cerebellum
Brainstem
Midbrain
Pons
Medulla oblongata
Spinal cord
Meningeal infiltration to CSF (p <0.001)
Cases
=1 case
=5 cases

Figure 5.11 *(preceding page)*: Phenotypic and tumor location distinctions of the multi-omic defined PCNSL subtypes. Panel A shows a heatmap with either gene signature activity (measured by GSVA) or immune cell proportions (CiberSortx deconvoluted) across molecular subtypes. P-values indicate higher expression of the colored group when compared against the others (Wilcoxon-test, left side of plot). FRC, Fibroblastic reticular cells; FDC, follicular dendritic cells. Panel B shows the tumor location of 90 PCNSLs (FFPE cohort) in the human central nervous system grouped by molecular subtype where the number of cases is indicated within the circles. Tumors occurring in midline locations are depicted in the sagittal view (left panel), meanwhile, tumors occurring in the cerebral and cerebellar hemispheres are depicted in the exterior view (right panel). P-value refers to a one-sided Fisher test (event within-cluster vs outside-cluster).


A
Hypermethylated
Hypomethylated
−Log10(p.adjusted)
CS group
CS1 (n=27)
CS2 (n=17)
CS3 (n=11)
CS4 (n=9)
Methylation direction
Hypermethylated
Hypomethylated
1.28
1.82
1.78
2.85
2.29
1.68
0.63
2.8
5.4
2.37
2.69
8
2.16
7.74
2.15
4.56
4.52
6.29
3.24
8
1.83
2.75
1.43
neurogenesis
cell-cell signaling
hemopoiesis
regulation of immune response
posttranscriptional gene silencing
regulation of B-cell activation
B-cell differentiation
regulation of lymphocyte proliferation
negative regulation of superoxide dismutase activity
regulation of cell-cell adhesion
regulation of cell communication
antigen receptor-mediated signaling pathway
positive regulation of cytokine production involved in inflammatory response
response to oxygen radical
nucleotide-excision repair, DNA damage recognition
regulation of response to oxidative stress
positive regulation of CD24 production
mature B-cell differentiation involved in immune response
positive regulation of leukocyte cell-cell adhesion
regulation of interleukin-1 production
positive regulation of immunological synapse formation
GO Term

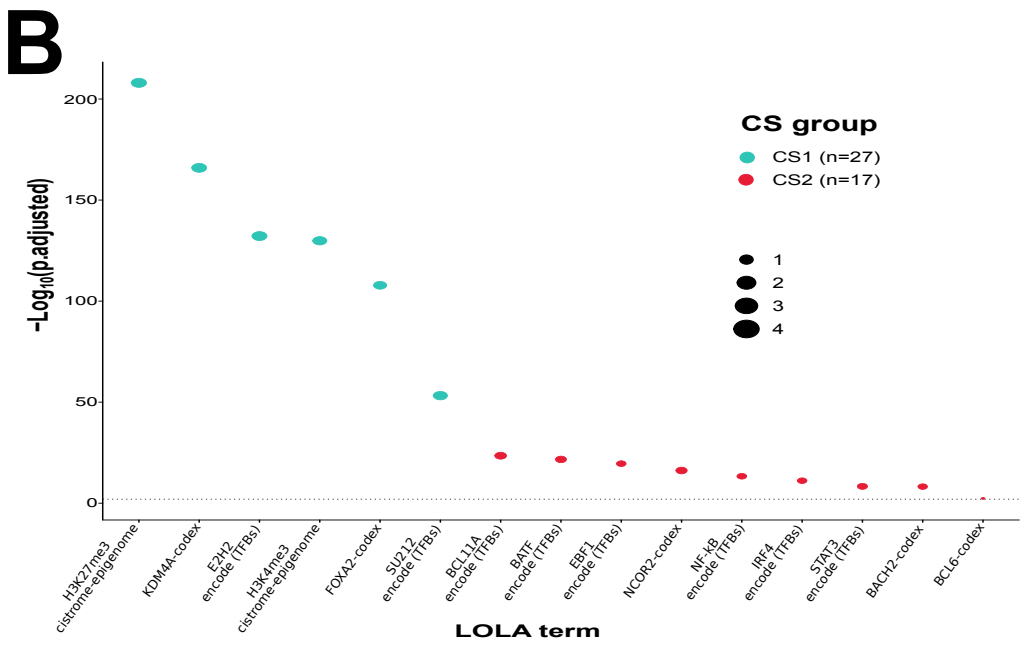

B
−Log10(p.adjusted)
CS group
CS1 (n=27)
CS2 (n=17)
1
2
3
4
H3K27me3 cistrome-epigenome
KDM4A-codex
E2H2 encode (TFBs)
H3K4me3 cistrome-epigenome
FOXA2-codex
SU212 encode (TFBs)
BCL11A encode (TFBs)
BATF encode (TFBs)
EBF1 encode (TFBs)
NCOR2-codex
NF-kB encode (TFBs)
IRF4 encode (TFBs)
STAT3 encode (TFBs)
BACH2-codex
BCL6-codex
LOLA term

Figure 5.12 *(preceding page)*: Epigenetic attributes across PCNSL subtypes. Panel A shows GO enrichment analysis on DMP across subtypes where the log(OddsRatio) is annotated next to its associated p-adjusted value. Panel B shows the locus overlap (LOLA) region set enrichment analysis for hypermethylated promoters across the four PCNSL subtypes. The CS3 group is characterized by enrichment of H3K27me3 probably resulting from the high PRC2 complex transcriptomic activity, whereas the CS2 by enrichment of BCL11A, NF-$\kappa\beta$, and IRF4 which is in line with the observed low transcriptomic activity of the related targets (Figure 3A). X-axis presents the targets followed by the database. P-values were calculated using a two-sided Fisher's exact test and then adjusted for multiple testing by the FDR method. The complete lists of GO and LOLA enrichments results are provided in Tables S12 and S13.

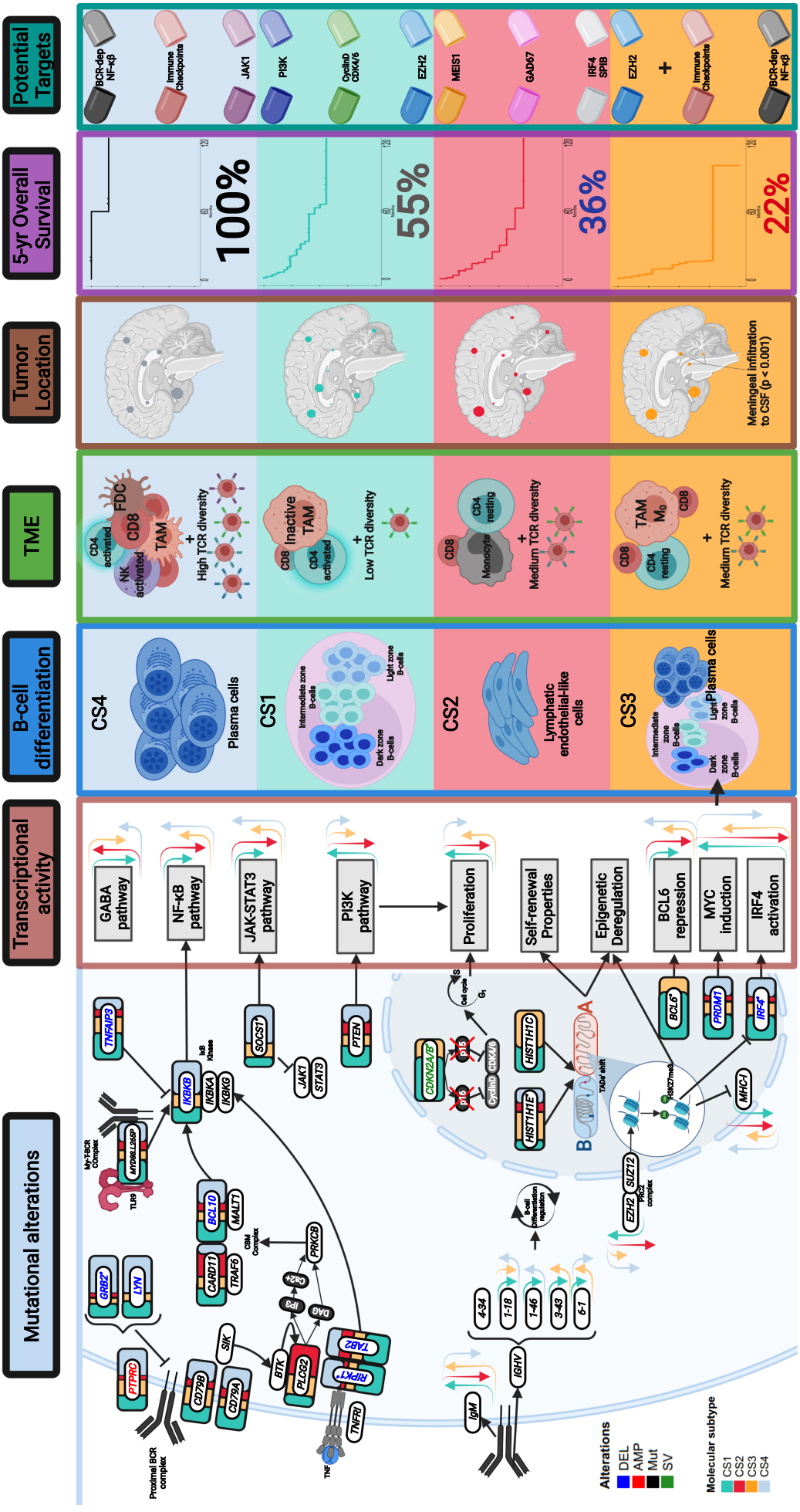

Potential Targets
5-yr Overall Survival
Tumor Location
TME
B-cell differentiation
Transcriptional activity
Mutational alterations
BCR-dep NF-κβ
Immune Checkpoints
JAK1
PI3K
CyclinD CDK4/6
EZH2
MEIS1
GAD67
IRF4 SPIB
EZH2
+
Immune Checkpoints
BCR-dep NF-κβ
100%
55%
36%
22%
Meningeal infiltration to CSF (p < 0.001)
CS4
Plasma cells
CS1
CS2
Lymphatic endothelial-like cells
CS3
Plasma cells
High TCR diversity
Low TCR diversity
Medium TCR diversity
Medium TCR diversity
GABA pathway
NF-κB pathway
JAK-STAT3 pathway
PI3K pathway
Proliferation
Self-renewal Properties
Epigenetic Deregulation
BCL6 repression
MYC induction
IRF4 activation
Alterations
DEL
AMP
Mut
SV
Molecular subtype
CS1
CS2
CS3
CS4

Figure 5.13 *(preceding page)*: From multi-omics to potential therapeutic targets. Shown is a schematic representation summarizing the major molecular findings and proposed potential therapeutic targets. Contribution of each molecular subtype to the indicated alteration where color bar width indicates the prevalence of each subtype (Mutational alterations section). Asterisks indicate if a genetic alteration is enriched in any CS subtype (related to Figure 2C). Arrows indicate transcriptional gene signature activity where the height indicates the relative up or downregulation (according to Figure 3A). DEL, deletion; AMP, amplification; Mut, mutation; SV refers to fusion transcripts. NF-$\kappa\beta$ activity could not be explained by self-antigen-dependent chronic active BCR signaling upregulation since $IgV_{H4\text{-}34}$ expression was not significantly different across groups. The IgVH regions more expressed in CS4 ($p < 0.05$) were the $V_{1\text{-}18}$, $V_{1\text{-}46}$, and $V_{3\text{-}9}$ that have been associated with more differentiated B-cell stages; meanwhile, the naive-transition stage-related, $V_{3\text{-}43}$ and $V_{4\text{-}30\text{-}2}$ regions, were upregulated in CS3. *HIST1H1E/C* mutations confer self-renewal properties to B-cells and induce shifts from compartment B to compartment A chromatin. Analyses for determining the TCR/BCR diversity, the immunoglobulin heavy-chain variable (V_H) and constant regions expression, and the master regulators (*MEIS1, IRF4, and SPIB*) listed as potential CS2 targets are provided in Supplementary Appendix 1.

Chapter 6

General discussion and conclusion

6.1 General discussion

The major focus of this thesis has been to identify and characterize molecular PCNSL subtypes with shared pathogenesis which will ultimately give plausible explanations to the PCNSL response heterogeneity. While Chapter 5.1 served to this major focus by covering HLA structure and diversity, which has been proved to be constantly disrupted (e.g., by deletions or mutations) in PCNSL (Chapuy et al., 2016, 2018; Schmitz et al., 2018); Chapter 5.2 helps by developing and validating a code to identify c-AID mutations which are fundamental in B-cell biology and B-cell lymphomagenesis (Chapuy et al., 2018).

HLA structure/diversity and genetic susceptibility in PCNSL and other B-cell NHLs

B-cell NHLs' (including PCNSL) risk associations were initially attributed to family history of the disease, inflammation, and immune components including HLA genetic variations; however, recent GWAS have broad more information into the subject. In Chapter 5.1, I review the HLA structure and its diversity and summarized all the original articles showing evidence of genetic variations on five NHL subtypes (DLBCL, FL, CLL, MZL, and PCNSL).

In the literature review article, we showed that the HLA variants are the most studied within the B-cell NHL context since that region is critical for innate and adaptive immune responses. Interestingly, HLA status has been proved to be a risk factor in B-cell NHL by promoting immune escape, this has also been observed specifically for PCNSL (Chapuy et al., 2016, 2018; Din et al., 2019; Moore et al., 2020; Schmitz et al., 2018; Zhong, Cozen, Bolanos, Song, & Wang, 2019). As reviewed in Chapter 2.3.2, antigens/neoantigens production is an important tumor escape mechanism from immune surveillance, this can be disrupted as a

consequence of HLA homozygosity as seen in most of the reviewed lymphomas (including PCNSL) (Mueller & Machulla, 2002; S. S. Wang et al., 2018).

Moreover, specifically speaking of PCNSL, the only study evaluating associations between genetic variants and PCNSL risk, was done by our group in a French cohort (Labreche et al., 2019). Though this study found some additional variants associated with PCNSL risk, it is clear that additional studies are needed to better elucidate PCNSL pathogenesis.

Implications of AID-related mutations at pan-cancer level

As reviewed throughout this thesis (see Chapters 2.1.1 and 2.1.2), AID off-target activity is in the context of B-cell biology and lymphomagenesis. In this study, which is the largest to date, I integrated more than 50,000 bulk level samples and 2.5 million cells at single-cell resolution across 80 tumor types (including B-cell malignancies) and different data levels. The main objective, within the article, was to thoroughly describe the oncogenic and clinical implications of AID off-target mutations at pan-cancer scale; however, the major goal of this section regarding this thesis was to develop and validate the code to target the c-AID mutations.

Firstly, we demonstrated that *AICDA* expression is only present in normal B-cells by using a series of single-cell RNA-seq studies; nevertheless, this changes after malignant transformation since we observed its expression across different cancer types at single-cell resolution. Next, we evaluated our code for tracking c-AID mutations using tetranucleotide motifs by firstly, applying it to a series of hematological cancers and finding already reported AID targets (e.g., *PIM1*, *HIST1H1C*), secondly, evaluating that our code does not identify the same mutations as other COSMIC somatic signatures, and finally, ruling out that the observed AID mutations were generated by chance. Moreover, we also described, as expected, that the frequency of c-AID mutations is higher in hematological cancer compared to others.

After validating the code, we described the landscape and implications of c-AID mutations. We found that c-AID activity occurs mainly during the transcription of its off-target genes and is increased in MSI tumors. Additionally, we showed that in some cancer types AID promiscuous activity aims for least-positive selection hotspots that synergize with previous stronger hotspot mutations (minor mutation PIK3CA E726, especially present in SKCM and BRCA). Finally, we demonstrated that the AID-related fraction of mutations is an independent prognostic value to ICI and presented different analyses to explain such findings.

The recompilation of all the public datasets along with the findings and the code to detect c-AID mutations, provide the basis for testing the potential role of c-AID mutations in hematological and non-hematological cancers. However, due to the bioinformatic nature of the study, several biological validations of the results

have to be performed.

Multi-omic data integration reveals PCNSL molecular subtypes with shared pathogenesis and clinical outcome implications

To the best of our knowledge, the study presented in Chapter 5.3 represents the largest multi-omic study of PCNSL conducted to date. Our study builds on the current classification of DLBCL, the MCD/C5 DLBCLs (see Chapter 2.2), by the addition of the molecular heterogeneity within PCNSL that may inform on its pathogenesis and ultimately give potential therapeutic targets. Here, I found four PCNSL subtypes with shared multi-omic features such as distinct oncogenic pathways, gene expression phenotypes, methylation profiles, TME, and clinico-radiological characteristics.

Our findings help to elucidate the highly heterogeneous response in PCNSL by connecting different multi-genomic layers with the clinicOmic information. Here we showed that the CS4 group shares a constitutive NF-$\kappa\beta$ activation, which is one of the main features of the MCD or C5 DLBCL subtypes (Chapuy et al., 2016, 2018; Schmitz et al., 2018; G. W. Wright et al., 2020), with the CS3 group, however, their clinical outcomes in both OS and PFS are totally opposite. We showed that such variations are mainly due to the more aggressive tumor locations for CS3 and a hot TME for CS4. Regarding the potential therapeutic targets, even though both groups could be potentially more sensitive to BTK inhibitors (e.g., ibrutinib), the CS4 group could also benefit from JAK1 and immune checkpoint inhibitors either because it presents high JAK-STAT transcriptional activity or high MHC-I expression with the absence HLA biallelic deletions. Moreover, the CS3 group could also benefit from ICI but only after exposure to EZH2 inhibitors since it could restore its missing MHC-I expression.

Interestingly, the CS1 and CS2 PCNSL subtypes were largely hypermethylated when compared to the others, which has been previously associated with a cold TME (Kotlov et al., 2021), as observed transcriptionally. For the CS1, the high PRC2 complex activity and proliferation (driven by genetic alterations involved in the cell cycle) were directly "seen" in its hypermethylator phenotype. On the other hand, the "disrupted" B-cell differentiation programs observed transcriptionally in CS2 were corroborated at the epigenetic level. We proposed that the CS1 immune cold group could be responsive to cyclin CDK4 and CDK6 plus PI3K inhibitors; while the CS2 group may be potentially susceptible to inhibition of the TFs IRF4 (e.g., lenalidomide), SPIB, and MEIS1 (e.g., MEISi-1), and/or inhibition of GAD67.

Regarding c-AID off-target activity, even though we did not observe any difference in the global number of c-AID mutations across the molecular subtypes, we showed that globally (using all the PCNSL cohort) both c-AID and non c-AID

(Cosmic signature SBS9) mutations occurs at early stages of PCNSL tumorigenesis, thus reflecting its importance for PCNSL pathogenesis.

Since acquiring FF tissue for PCNSL is not routinely performed in the clinics, we validated our results in an additional FFPE cohort. Additionally, we developed RBraLymP (RNA-based Brain Lymphoma Profiler), which uses gene expression data from either FFPE or FF tissue, to identify the PCNSL molecular subtypes associated with multi-omic features. We made the code publicly accessible to incentive researchers around the world directing new therapy efforts to the most appropriate PCNSL patients.

6.2 General conclusion

The understanding of the molecular and clinical response heterogeneity in PCNSL had not been properly addressed since it was built on the current classification of DLBCL which consisted of a low number of PCNSL samples. The collective findings of my thesis amend this gap by linking the integrated multi-omic features within each molecular PCNSL subtype to potential treatment targets. Moreover, the RNA-based algorithm, RBraLymP, can facilitate future efforts for developing and evaluating targeted therapeutic approaches for these poorly understood and highly deadly malignancies.

Finally, during my thesis, I contributed to the bioinformatic analysis on different glioblastoma studies which are listed in APPENDIX 2.

References

10 Abrey, L. E., Ben-Porat, L., Panageas, K. S., Yahalom, J., Berkey, B., Curran, W., ... DeAngelis, L. M. (2006). Primary Central Nervous System Lymphoma: The Memorial Sloan-Kettering Cancer Center Prognostic Model. *Journal of Clinical Oncology*, *24*(36), 5711–5715. http://doi.org/10.1200/JCO.2006.08.2941

Akdemir, K. C., Le, V. T., Kim, J. M., Killcoyne, S., King, D. A., Lin, Y.-P., ... Andrew Futreal, P. (2020). Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nature Genetics*, *52*(11), 1178–1188. http://doi.org/10.1038/s41588-020-0708-0

Alame, M., Cornillot, E., Cacheux, V., Rigau, V., Costes-Martineau, V., Lacheretz-Szablewski, V., & Colinge, J. (2021). The immune contexture of primary central nervous system diffuse large B cell lymphoma associates with patient survival and specific cell signaling. *Theranostics*, *11*(8), 3565–3579. http://doi.org/10.7150/thno.54343

Alcantara, M., Fuentealba, J., & Soussain, C. (2021). Emerging Landscape of Immunotherapy for Primary Central Nervous System Lymphoma. *Cancers*, *13*(20), 5061. http://doi.org/10.3390/cancers13205061

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., ... Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503–511. http://doi.org/10.1038/35000501

Alvarez, M. J., Shen, Y., Giorgi, F. M., Lachmann, A., Ding, B. B., Ye, B. H., & Califano, A. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics*, *48*(8), 838–847. http://doi.org/10.1038/ng.3593

Amiel, J. (1967). Study of the Leukocyte Phenotypes in Hodgkin's Disease. In *Histocompatibility Testing* (Teraski, P. I., pp. 79–81). Munksgaard; Copenhagen.

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, *22*(10), 2008–2017. http://doi.org/10.1101/gr.133744.111

Andrews, S. (2010). *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.* Retrieved from

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arthur, S. E., Jiang, A., Grande, B. M., Alcaide, M., Cojocaru, R., Rushton, C. K., ... Morin, R. D. (2018). Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nature Communications*, *9*(1), 4001. http://doi.org/10.1038/s41467-018-06354-3

Aslan, K., Turco, V., Blobner, J., Sonner, J. K., Liuzzi, A. R., Núñez, N. G., ... Platten, M. (2020). Heterogeneity of response to immune checkpoint blockade in hypermutated experimental gliomas. *Nature Communications*, *11*(1), 931. http://doi.org/10.1038/s41467-020-14642-0

Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Alexandrov, L. B., Nik-Zainal, S., ... Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. http://doi.org/10.1038/nature12477

Álvarez-Prado, Á. F., Pérez-Durán, P., Pérez-García, A., Benguria, A., Torroja, C., Yébenes, V. G. de, & Ramiro, A. R. (2018). A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. *Journal of Experimental Medicine*, *215*(3), 761–771. http://doi.org/10.1084/jem.20171738

Baecklund, F., Foo, J.-N., Bracci, P., Darabi, H., Karlsson, R., Hjalgrim, H., ... Smedby, K. E. (2014). A comprehensive evaluation of the role of genetic variation in follicular lymphoma survival. *BMC Medical Genetics*, *15*(1), 113. http://doi.org/10.1186/s12881-014-0113-6

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., ... Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, *462*(7269), 108–112. http://doi.org/10.1038/nature08460

Barrios, O. de, Meler, A., & Parra, M. (2020). MYC's Fine Line Between B Cell Development and Malignancy. *Cells*, *9*(2), 523. http://doi.org/10.3390/cells9020523

Bassig, B. A., Cerhan, J. R., Au, W.-Y., Kim, H. N., Sangrajrang, S., Hu, W., ... Rothman, N. (2015). Genetic susceptibility to diffuse large B-cell lymphoma in a pooled study of three Eastern Asian populations. *European Journal of Haematology*, *95*(5), 442–448. http://doi.org/10.1111/ejh.12513

Basso, K., & Dalla-Favera, R. (2010). BCL6: Master regulator of the germinal center reaction and key oncogene in B cell lymphomagenesis. *Advances in Immunology*, *105*, 193–210. http://doi.org/10.1016/S0065-2776(10)05007-8

Baumgarten, L. von, Illerhaus, G., Korfel, A., Schlegel, U., Deckert, M., & Dreyling, M. (2018). The Diagnosis and Treatment of Primary CNS Lymphoma. *Dtsch Arztebl International*, *115*(25), 419–426. http://doi.org/10.3238/arztebl.2018.0419

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., ... Reyniès, A. de. (2016). Estimating the population abundance of tissue-

infiltrating immune and stromal cell populations using gene expression. *Genome Biology*, *17*(1), 218. http://doi.org/10.1186/s13059-016-1070-5

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models: GENERATING SURVIVAL TIMES. *Statistics in Medicine*, *24*(11), 1713–1723. http://doi.org/10.1002/sim.2059

Bergstrom, E. N., Barnes, M., Martincorena, I., & Alexandrov, L. B. (2020). Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinformatics*, *21*(1), 438. http://doi.org/10.1186/s12859-020-03772-3

Bernatsky, S., Velásquez García, H. A., Spinelli, J. J., Gaffney, P., Smedby, K. E., Ramsey-Goldman, R., … Clarke, A. E. (2017). Lupus-related single nucleotide polymorphisms and risk of diffuse large B-cell lymphoma. *Lupus Science & Medicine*, *4*(1), e000187. http://doi.org/10.1136/lupus-2016-000187

Berndt, Sonja I., Camp, N. J., Skibola, C. F., Vijai, J., Wang, Z., Gu, J., … Slager, S. L. (2016). Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nature Communications*, *7*(1), 10933. http://doi.org/10.1038/ncomms10933

Berndt, Sonja I., Skibola, C. F., Joseph, V., Camp, N. J., Nieters, A., Wang, Z., … Slager, S. L. (2013). Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nature Genetics*, *45*(8), 868–876. http://doi.org/10.1038/ng.2652

Boegel, S., Castle, J. C., Kodysh, J., O'Donnell, T., & Rubinsteyn, A. (2019). Bioinformatic methods for cancer neoantigen prediction. In *Progress in Molecular Biology and Translational Science* (Vol. 164, pp. 25–60). Elsevier. http://doi.org/10.1016/bs.pmbts.2019.06.016

Boichard, A., Pham, T. V., Yeerna, H., Goodman, A., Tamayo, P., Lippman, S., … Kurzrock, R. (2018). APOBEC-related mutagenesis and neo-peptide hydrophobicity: Implications for response to immunotherapy. *Oncoimmunology*, *8*(3). http://doi.org/10.1080/2162402X.2018.1550341

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. http://doi.org/10.1093/bioinformatics/btu170

Bolotin, D. A., Poslavsky, S., Davydov, A. N., Frenkel, F. E., Fanchi, L., Zolotareva, O. I., … Chudakov, D. M. (2017). Antigen receptor repertoire profiling from RNA-seq data. *Nature Biotechnology*, *35*(10), 908–911. http://doi.org/10.1038/nbt.3979

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., & Chudakov, D. M. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. *Nature Methods*, *12*(5), 380–381. http://doi.org/10.1038/nmeth.3364

Bowzyk Al-Naeeb, A., Ajithkumar, T., Behan, S., & Hodson, D. J. (2018). Non-Hodgkin lymphoma. *BMJ*, k3204. http://doi.org/10.1136/bmj.k3204

Brabletz, T., Pfeuffer, I., Schorr, E., Siebelt, F., Wirth, T., & Serfling, E. (1993). Transforming growth factor beta and cyclosporin A inhibit the inducible activity of the interleukin-2 gene in T cells through a noncanonical octamer-binding site. *Molecular and Cellular Biology*, *13*(2), 1155–1162. http://doi.org/10.1128/mcb.13.2.1155-1162.1993

Bracci, P. M., Benavente, Y., Turner, J. J., Paltiel, O., Slager, S. L., Vajdic, C. M., ... Sanjose, S. de. (2014). Medical History, Lifestyle, Family History, and Occupational Risk Factors for Marginal Zone Lymphoma: The InterLymph Non-Hodgkin Lymphoma Subtypes Project. *JNCI Monographs*, *2014*(48), 52–65. http://doi.org/10.1093/jncimonographs/lgu011

Braggio, E., Van Wier, S., Ojha, J., McPhail, E., Asmann, Y. W., Egan, J., ... O'Neill, B. P. (2015). Genome-Wide Analysis Uncovers Novel Recurrent Alterations in Primary Central Nervous System Lymphomas. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *21*(17), 3986–3994. http://doi.org/10.1158/1078-0432.CCR-14-2116

Branton, S. A., Ghorbani, A., Bolt, B. N., Fifield, H., Berghuis, L. M., & Larijani, M. (2020). Activation-induced cytidine deaminase can target multiple topologies of double-stranded DNA in a transcription-independent manner. *The FASEB Journal*, *34*(7), 9245–9268. http://doi.org/10.1096/fj.201903036RR

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. http://doi.org/10.1038/nbt.3519

Broekman, M. L., Maas, S. L. N., Abels, E. R., Mempel, T. R., Krichevsky, A. M., & Breakefield, X. O. (2018). Multidimensional communication in the microenvirons of glioblastoma. *Nature Reviews Neurology*, *14*(8), 482–495. http://doi.org/10.1038/s41582-018-0025-8

Bruno, A., Boisselier, B., Labreche, K., Marie, Y., Polivka, M., Jouvet, A., ... Hoang-Xuan, K. (2014). Mutational analysis of primary central nervous system lymphoma. *Oncotarget*, *5*(13). http://doi.org/10.18632/oncotarget.2080

Cambruzzi, E. (2020). Primary Intra-Axial Diffuse Large B-Cell Lymphoma in Immunocompetent Patients: Clinical Impact of Molecular Analysis and Histogenetic Evaluation. *World Neurosurgery*, *134*, 215–220. http://doi.org/10.1016/j.wneu.2019.09.158

Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, *10*(1), 3240. http://doi.org/10.1038/s41467-019-11146-4

Camilleri-Broët, S., Martin, A., Moreau, A., Angonin, R., Hénin, D.,

Gontier, M. F., ... Raphaël, M. (1998). Primary Central Nervous System Lymphomas in 72 Immunocompetent Patients: Pathologic Findings and Clinical Correlations. *American Journal of Clinical Pathology*, *110*(5), 607–612. http://doi.org/10.1093/ajcp/110.5.607

Cannataro, V. L., Gaffney, S. G., Sasaki, T., Issaeva, N., Grewal, N. K. S., Grandis, J. R., ... Townsend, J. P. (2019). APOBEC-induced mutations and their cancer effect size in head and neck squamous cell carcinoma. *Oncogene*, *38*(18), 3475–3487. http://doi.org/10.1038/s41388-018-0657-6

Carbone, A., Roulland, S., Gloghini, A., Younes, A., Keudell, G. von, López-Guillermo, A., & Fitzgibbon, J. (2019). Follicular lymphoma. *Nature Reviews Disease Primers*, *5*(1), 83. http://doi.org/10.1038/s41572-019-0132-x

Carrassa, L., Colombo, I., Damia, G., & Bertoni, F. (2020). Targeting the DNA damage response for patients with lymphoma: Preclinical and clinical evidences. *Cancer Treatment Reviews*, *90*, 102090. http://doi.org/10.1016/j.ctrv.2020.102090

Casellas, R., Basu, U., Yewdell, W. T., Chaudhuri, J., Robbiani, D. F., & Di Noia, J. M. (2016). Mutations, kataegis and translocations in B cells: Understanding AID promiscuous activity. *Nature Reviews Immunology*, *16*(3), 164–176. http://doi.org/10.1038/nri.2016.2

Castro, M. A. A., Santiago, I. de, Campbell, T. M., Vaughn, C., Hickey, T. E., Ross, E., ... Meyer, K. B. (2016). Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nature Genetics*, *48*(1), 12–21. http://doi.org/10.1038/ng.3458

Castro, M. A., Wang, X., Fletcher, M. N., Meyer, K. B., & Markowetz, F. (2012). RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. *Genome Biology*, *13*(4), R29. http://doi.org/10.1186/gb-2012-13-4-r29

Cerhan, James R., Berndt, S. I., Vijai, J., Ghesquières, H., McKay, J., Wang, S. S., ... Chanock, S. J. (2014). Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nature Genetics*, *46*(11), 1233–1238. http://doi.org/10.1038/ng.3105

Cerhan, James R., Fredericksen, Z. S., Novak, A. J., Ansell, S. M., Kay, N. E., Liebow, M., ... Slager, S. L. (2012). A two-stage evaluation of genetic variation in immune and inflammation genes with risk of non-Hodgkin lymphoma identifies new susceptibility locus in 6p21.3 region. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *21*(10), 1799–1806. http://doi.org/10.1158/1055-9965.EPI-12-0696

Cerhan, J. R., Kricker, A., Paltiel, O., Flowers, C. R., Wang, S. S., Monnereau, A., ... Skibola, C. F. (2014). Medical History, Lifestyle, Family History, and Occupational Risk Factors for Diffuse Large B-Cell Lymphoma: The InterLymph

Non-Hodgkin Lymphoma Subtypes Project. *JNCI Monographs*, *2014*(48), 15–25. http://doi.org/10.1093/jncimonographs/lgu010

Chalise, P., & Fridley, B. L. (2017). Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLOS ONE*, *12*(5), e0176278. http://doi.org/10.1371/journal.pone.0176278

Chapuy, B., Roemer, M. G. M., Stewart, C., Tan, Y., Abo, R. P., Zhang, L., ... Shipp, M. A. (2016). Targetable genetic features of primary testicular and primary central nervous system lymphomas. *Blood*, *127*(7), 869–881. http://doi.org/10.1182/blood-2015-10-673236

Chapuy, B., Stewart, C., Dunford, A. J., Kim, J., Kamburov, A., Redd, R. A., ... Shipp, M. A. (2018). Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nature Medicine*, *24*(5), 679–690. http://doi.org/10.1038/s41591-018-0016-8

Chen, C., Zhuo, H., Wei, X., & Ma, X. (2019). Contrast-Enhanced MRI Texture Parameters as Potential Prognostic Factors for Primary Central Nervous System Lymphoma Patients Receiving High-Dose Methotrexate-Based Chemotherapy. *Contrast Media & Molecular Imaging*, *2019*, 1–7. http://doi.org/10.1155/2019/5481491

Chen, D. S., & Mellman, I. (2013). Oncology Meets Immunology: The Cancer-Immunity Cycle. *Immunity*, *39*(1), 1–10. http://doi.org/10.1016/j.immuni.2013.07.012

Chin, L., Hahn, W. C., Getz, G., & Meyerson, M. (2011). Making sense of cancer genomic data. *Genes & Development*, *25*(6), 534–555. http://doi.org/10.1101/gad.2017311

Cho, H., Kim, S. H., Kim, S.-J., Chang, J. H., Yang, W. I., Suh, C.-O., ... Kim, J. S. (2017). The prognostic role of CD68 and FoxP3 expression in patients with primary central nervous system lymphoma. *Annals of Hematology*, *96*(7), 1163–1173. http://doi.org/10.1007/s00277-017-3014-x

Cho, I., Lee, H., Yoon, S. E., Ryu, K. J., Ko, Y. H., Kim, W. S., & Kim, S. J. (2020). Serum levels of soluble programmed death-ligand 1 (sPD-L1) in patients with primary central nervous system diffuse large B-cell lymphoma. *BMC Cancer*, *20*(1), 120. http://doi.org/10.1186/s12885-020-6612-2

Chunsong, H., Yuling, H., Li, W., Jie, X., Gang, Z., Qiuping, Z., ... Jinquan, T. (2006). CXC Chemokine Ligand 13 and CC Chemokine Ligand 19 Cooperatively Render Resistance to Apoptosis in B Cell Lineage Acute and Chronic Lymphocytic Leukemia CD23 $^{+}$ CD5 $^{+}$ B Cells. *The Journal of Immunology*, *177*(10), 6713–6722. http://doi.org/10.4049/jimmunol.177.10.6713

Conde, L., Halperin, E., Akers, N. K., Brown, K. M., Smedby, K. E., Rothman, N., ... Skibola, C. F. (2010). Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nature Genetics*, *42*(8), 661–664. http://doi.org/10.1038/ng.626

Crowther-Swanepoel, D., Broderick, P., Di Bernardo, M. C., Dobbins, S. E., Torres, M., Mansouri, M., ... Houlston, R. S. (2010). Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nature Genetics*, *42*(2), 132–136. http://doi.org/10.1038/ng.510

Crowther-Swanepoel, D., Mansouri, M., Enjuanes, A., Vega, A., Smedby, K. E., Ruiz-Ponte, C., ... Houlston, R. S. (2010). Verification that common variation at 2q37.1, 6p25.3, 11q24.1, 15q23, and 19q13.32 influences chronic lymphocytic leukaemia risk: Short Report. *British Journal of Haematology*, no–no. http://doi.org/10.1111/j.1365-2141.2010.08270.x

Deakin, J. E., Papenfuss, A. T., Belov, K., Cross, J. G., Coggill, P., Palmer, S., ... Graves, J. A. M. (2006). Evolution and comparative analysis of the MHC Class III inflammatory region. *BMC Genomics*, *7*(1), 281. http://doi.org/10.1186/1471-2164-7-281

Deckert, M., Engert, A., Brück, W., Ferreri, A. J. M., Finke, J., Illerhaus, G., ... DeAngelis, L. M. (2011). Modern concepts in the biology, diagnosis, differential diagnosis and treatment of primary central nervous system lymphoma. *Leukemia*, *25*(12), 1797–1807. http://doi.org/10.1038/leu.2011.169

Deckert, Martina, Montesinos-Rongen, M., Brunn, A., & Siebert, R. (2014). Systems biology of primary CNS lymphoma: From genetic aberrations to modeling in mice. *Acta Neuropathologica*, *127*(2), 175–188. http://doi.org/10.1007/s00401-013-1202-x

Delgado, P., Álvarez-Prado, Á. F., Marina-Zárate, E., Sernandez, I. V., Mur, S. M., Barrera, J. de la, ... Ramiro, A. R. (2020). Interplay between UNG and AID governs intratumoral heterogeneity in mature B cell lymphoma. *PLOS Genetics*, *16*(12), e1008960. http://doi.org/10.1371/journal.pgen.1008960

Dendrou, C. A., Petersen, J., Rossjohn, J., & Fugger, L. (2018). HLA variation and disease. *Nature Reviews Immunology*, *18*(5), 325–339. http://doi.org/10.1038/nri.2017.143

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. http://doi.org/10.1038/ng.806

Dersh, D., Phelan, J. D., Gumina, M. E., Wang, B., Arbuckle, J. H., Holly, J., ... Yewdell, J. W. (2021). Genome-wide Screens Identify Lineage- and Tumor-Specific Genes Modulating MHC-I- and MHC-II-Restricted Immunosurveillance of Human Lymphomas. *Immunity*, *54*(1), 116–131.e10. http://doi.org/10.1016/j.immuni.2020.11.002

Di Bernardo, M. C., Crowther-Swanepoel, D., Broderick, P., Webb, E., Sellick, G., Wild, R., ... Houlston, R. S. (2008). A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nature Genetics*,

40(10), 1204–1210. http://doi.org/10.1038/ng.219

Di Paolo, A., Arrigoni, E., Luci, G., Cucchiara, F., Danesi, R., & Galimberti, S. (2019). Precision Medicine in Lymphoma by Innovative Instrumental Platforms. *Frontiers in Oncology*, *9*, 1417. http://doi.org/10.3389/fonc.2019.01417

Din, L., Sheikh, M., Kosaraju, N., Smedby, K. E., Bernatsky, S., Berndt, S. I., … Khankhanian, P. (2019). Genetic overlap between autoimmune diseases and non-Hodgkin lymphoma subtypes. *Genetic Epidemiology*, *43*(7), 844–863. http://doi.org/10.1002/gepi.22242

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. http://doi.org/10.1093/bioinformatics/bts635

Dorsett, Y., McBride, K. M., Jankovic, M., Gazumyan, A., Thai, T.-H., Robbiani, D. F., … Nussenzweig, M. C. (2008). MicroRNA-155 Suppresses Activation-Induced Cytidine Deaminase-Mediated Myc-Igh Translocation. *Immunity*, *28*(5), 630–638. http://doi.org/10.1016/j.immuni.2008.04.002

Duclos, C. M., Champagne, A., Carrier, J. C., Saucier, C., Lavoie, C. L., & Denault, J.-B. (2017). Caspase-mediated proteolysis of the sorting nexin 2 disrupts retromer assembly and potentiates Met/hepatocyte growth factor receptor signaling. *Cell Death Discovery*, *3*(1), 16100. http://doi.org/10.1038/cddiscovery.2016.100

Dudley, W. N., Wickham, R., & Coombs, N. (2016). An Introduction to Survival Statistics: Kaplan-Meier Analysis. *Journal of the Advanced Practitioner in Oncology*, *7*(1), 91–100. http://doi.org/10.6004/jadpro.2016.7.1.8

Duran-Ferrer, M., Clot, G., Nadeu, F., Beekman, R., Baumann, T., Nordlund, J., … Martín-Subero, J. I. (2020). The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. *Nature Cancer*, *1*(11), 1066–1081. http://doi.org/10.1038/s43018-020-00131-2

Ekström-Smedby, K. (2006). Epidemiology and etiology of non-Hodgkin lymphoma – a review. *Acta Oncologica*, *45*(3), 258–271. http://doi.org/10.1080/02841860500531682

Eloranta, S., Brånvall, E., Celsing, F., Papworth, K., Ljungqvist, M., Enblad, G., & Ekström-Smedby, K. (2018). Increasing incidence of primary central nervous system lymphoma but no improvement in survival in Sweden 2000-2013. *European Journal of Haematology*, *100*(1), 61–68. http://doi.org/10.1111/ejh.12980

Endo, Y., Marusawa, H., Kinoshita, K., Morisawa, T., Sakurai, T., Okazaki, I.-M., … Chiba, T. (2007). Expression of activation-induced cytidine deaminase in human hepatocytes via NF-κB signaling. *Oncogene*, *26*(38), 5587–5595. http://doi.org/10.1038/sj.onc.1210344

Fang, L. T., Afshar, P. T., Chhibber, A., Mohiyuddin, M., Fan, Y., Mu, J. C., … Lam, H. Y. K. (2015). An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biology*, *16*(1), 197.

http://doi.org/10.1186/s13059-015-0758-2

Fangazio, M., Ladewig, E., Gomez, K., Garcia-Ibanez, L., Kumar, R., Teruya-Feldstein, J., ... Dalla-Favera, R. (2021). Genetic mechanisms of HLA-I loss and immune escape in diffuse large B cell lymphoma. *Proceedings of the National Academy of Sciences*, *118*(22), e2104504118. http://doi.org/10.1073/pnas.2104504118

Farrall, A. L., & Smith, J. R. (2021). Changing Incidence and Survival of Primary Central Nervous System Lymphoma in Australia: A 33-Year National Population-Based Study. *Cancers*, *13*(3), 403. http://doi.org/10.3390/cancers13030403

Feig, C., Jones, J. O., Kraman, M., Wells, R. J. B., Deonarine, A., Chan, D. S., ... Fearon, D. T. (2013). Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer. *Proceedings of the National Academy of Sciences*, *110*(50), 20212–20217. http://doi.org/10.1073/pnas.1320318110

Fischer, A., Vázquez-García, I., Illingworth, C. J. R., & Mustonen, V. (2014). High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports*, *7*(5), 1740–1752. http://doi.org/10.1016/j.celrep.2014.04.055

Fletcher, M. N. C., Castro, M. A. A., Wang, X., Santiago, I. de, O'Reilly, M., Chin, S.-F., ... Meyer, K. B. (2013). Master regulators of FGFR2 signalling and breast cancer risk. *Nature Communications*, *4*(1), 2464. http://doi.org/10.1038/ncomms3464

Fong, L. W. R., Yang, D. C., & Chen, C.-H. (2017). Myristoylated alanine-rich C kinase substrate (MARCKS): A multirole signaling protein in cancers. *Cancer and Metastasis Reviews*, *36*(4), 737–747. http://doi.org/10.1007/s10555-017-9709-6

Four, M., Cacheux, V., Tempier, A., Platero, D., Fabbro, M., Marin, G., ... Szablewski, V. (2017). PD1 and PDL1 expression in primary central nervous system diffuse large B-cell lymphoma are frequent and expression of PD1 predicts poor survival: PD1 expression in lymphoma. *Hematological Oncology*, *35*(4), 487–496. http://doi.org/10.1002/hon.2375

Franca, R. A., Travaglino, A., Varricchio, S., Russo, D., Picardi, M., Pane, F., ... Mascolo, M. (2020). HIV prevalence in primary central nervous system lymphoma: A systematic review and meta-analysis. *Pathology - Research and Practice*, *216*(11), 153192. http://doi.org/10.1016/j.prp.2020.153192

Fukumura, K., Kawazu, M., Kojima, S., Ueno, T., Sai, E., Soda, M., ... Mano, H. (2016). Genomic characterization of primary central nervous system lymphoma. *Acta Neuropathologica*, *131*(6), 865–875. http://doi.org/10.1007/s00401-016-1536-2

Gandhi, M. K., Hoang, T., Law, S. C., Brosda, S., O'Rourke, K., Tobin, J. W. D., ... Keane, C. (2021). EBV-associated primary CNS lymphoma occurring

after immunosuppression is a distinct immunobiological entity. *Blood*, *137*(11), 1468–1477. http://doi.org/10.1182/blood.2020008520

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., & Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, *29*(8), 1363–1375. http://doi.org/10.1101/gr.240663.118

Garcilazo-Reyes, Y., Ibáñez-Juliá, M.-J., Hernández-Verdin, I., Nguyen-Them, L., Younan, N., Houillier, C., … Alentorn, A. (2020). Treating central nervous system lymphoma in the era of precision medicine. *Expert Review of Precision Medicine and Drug Development*, *5*(4), 275–281. http://doi.org/10.1080/23808993.2020.1777853

Georgiou, K., Chen, L., Berglund, M., Ren, W., Miranda, N. F. C. C. de, Lisboa, S., … Pan-Hammarström, Q. (2016). Genetic basis of PD-L1 overexpression in diffuse large B-cell lymphomas. *Blood*, *127*(24), 3026–3034. http://doi.org/10.1182/blood-2015-12-686550

Ghesquieres, H., Slager, S. L., Jardin, F., Veron, A. S., Asmann, Y. W., Maurer, M. J., … Cerhan, J. R. (2015). Genome-Wide Association Study of Event-Free Survival in Diffuse Large B-Cell Lymphoma Treated With Immunochemotherapy. *Journal of Clinical Oncology*, *33*(33), 3930–3937. http://doi.org/10.1200/JCO.2014.60.2573

Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I., & Chédin, F. (2012). R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Molecular Cell*, *45*(6), 814–825. http://doi.org/10.1016/j.molcel.2012.01.017

Goeman, J. J. (2009). L_1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical Journal*, NA–NA. http://doi.org/10.1002/bimj.200900028

Gorelick, A. N., Sánchez-Rivera, F. J., Cai, Y., Bielski, C. M., Biederstedt, E., Jonsson, P., … Taylor, B. S. (2020). Phase and context shape the function of composite oncogenic mutations. *Nature*, *582*(7810), 100–103. http://doi.org/10.1038/s41586-020-2315-8

Gragert, L., Fingerson, S., Albrecht, M., Maiers, M., Kalaycio, M., & Hill, B. T. (2014). Fine-mapping of HLA associations with chronic lymphocytic leukemia in US populations. *Blood*, *124*(17), 2657–2665. http://doi.org/10.1182/blood-2014-02-558767

Groeneveld, C. S., Chagas, V. S., Jones, S. J. M., Robertson, A. G., Ponder, B. A. J., Meyer, K. B., & Castro, M. A. A. (2019). RTNsurvival: An R/Bioconductor package for regulatory network survival analysis. *Bioinformatics*, *35*(21), 4488–4489. http://doi.org/10.1093/bioinformatics/btz229

Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N., & Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de

novo fusion transcript assembly-based methods. *Genome Biology*, *20*(1), 213. http://doi.org/10.1186/s13059-019-1842-9

Haldorsen, I. S., Krossnes, B. K., Aarseth, J. H., Scheie, D., Johannesen, T. B., Mella, O., & Espeland, A. (2007). Increasing incidence and continued dismal outcome of primary central nervous system lymphoma in Norway 1989–2003: Time trends in a 15-year national survey. *Cancer*, *110*(8), 1803–1814. http://doi.org/10.1002/cncr.22989

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., … Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, *172*(5), 1091–1107.e17. http://doi.org/10.1016/j.cell.2018.02.001

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., … Guo, G. (2020). Construction of a human cell landscape at single-cell level. *Nature*, *581*(7808), 303–309. http://doi.org/10.1038/s41586-020-2157-4

Hans, C. P. (2004). Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, *103*(1), 275–282. http://doi.org/10.1182/blood-2003-05-1545

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 83. http://doi.org/10.1186/s13059-017-1215-1

Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, *14*(1), 7. http://doi.org/10.1186/1471-2105-14-7

He, M., Zuo, C., Wang, J., Liu, J., Jiao, B., Zheng, J., & Cai, Z. (2013). Prognostic significance of the aggregative perivascular growth pattern of tumor cells in primary central nervous system diffuse large B-cell lymphoma. *Neuro-Oncology*, *15*(6), 727–734. http://doi.org/10.1093/neuonc/not012

Hess, L. M., Brnabic, A., Mason, O., Lee, P., & Barker, S. (2019). Relationship between Progression-free Survival and Overall Survival in Randomized Clinical Trials of Targeted and Biologic Agents in Oncology. *Journal of Cancer*, *10*(16), 3717–3727. http://doi.org/10.7150/jca.32205

Hjalgrim, H., Rostgaard, K., Johnson, P. C. D., Lake, A., Shield, L., Little, A.-M., … Jarrett, R. F. (2010). HLA-A alleles and infectious mononucleosis suggest a critical role for cytotoxic T-cell response in EBV-related Hodgkin lymphoma. *Proceedings of the National Academy of Sciences*, *107*(14), 6400–6405. http://doi.org/10.1073/pnas.0915054107

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., … Stuart, J. M. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, *158*(4), 929–944. http://doi.org/10.1016/j.cell.2014.06.049

Hoang-Xuan, K., Bessell, E., Bromberg, J., Hottinger, A. F., Preusser, M., Rudà, R., … Weller, M. (2015). Diagnosis and treatment of primary

CNS lymphoma in immunocompetent patients: Guidelines from the European Association for Neuro-Oncology. *The Lancet Oncology*, *16*(7), e322–e332. http://doi.org/10.1016/S1470-2045(15)00076-5

Holland, C. H., Szalai, B., & Saez-Rodriguez, J. (2020). Transfer of regulatory knowledge from human to mouse for functional genomics analysis. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1863*(6), 194431. http://doi.org/10.1016/j.bbagrm.2019.194431

Holmes, A. B., Corinaldesi, C., Shen, Q., Kumar, R., Compagno, N., Wang, Z., ... Basso, K. (2020). Single-cell analysis of germinal-center B cells informs on lymphoma cell of origin and outcome. *Journal of Experimental Medicine*, *217*(10), e20200483. http://doi.org/10.1084/jem.20200483

Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., ... Zhang, H. (2020). RNA sequencing: New technologies and applications in cancer research. *Journal of Hematology & Oncology*, *13*(1), 166. http://doi.org/10.1186/s13045-020-01005-x

Honjo, T., Kinoshita, K., & Muramatsu, M. (2002). Molecular Mechanism of Class Switch Recombination: Linkage with Somatic Hypermutation. *Annual Review of Immunology*, *20*(1), 165–196. http://doi.org/10.1146/annurev.immunol.20.090501.112049

Honjo, T., Muramatsu, M., & Fagarasan, S. (2004). Aid. *Immunity*, *20*(6), 659–668. http://doi.org/10.1016/j.immuni.2004.05.011

Horvat, M., Zadnik, V., Južnič Šetina, T., Boltežar, L., Pahole Goličnik, J., Novaković, S., & Jezeršek Novaković, B. (2018). Diffuse large B-cell lymphoma: 10 years' real-world clinical experience with rituximab plus cyclophosphamide, doxorubicin, vincristine and prednisolone. *Oncology Letters*, *15*(3), 3602–3609. http://doi.org/10.3892/ol.2018.7774

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10), R115. http://doi.org/10.1186/gb-2013-14-10-r115

Houillier, C., Soussain, C., Ghesquières, H., Soubeyran, P., Chinot, O., Taillandier, L., ... Gyan, E. (2020). Management and outcome of primary CNS lymphoma in the modern era: An LOC network study. *Neurology*, *94*(10), e1027–e1039. http://doi.org/10.1212/WNL.0000000000008900

Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics*, *8*, 84. http://doi.org/10.3389/fgene.2017.00084

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., ... Lo, R. S. (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*, *165*(1), 35–44. http://doi.org/10.1016/j.cell.2016.02.065

Hundal, J., Kiwala, S., McMichael, J., Miller, C. A., Xia, H., Wollam, A. T., ... Griffith, M. (2020). pVACtools: A computational toolkit to identify

and visualize cancer neoantigens. *Cancer Immunology Research*, *8*(3), 409–420. http://doi.org/10.1158/2326-6066.CIR-19-0401

ImmunoMind Team. (2019). *Immunarch: An R Package for Painless Analysis of Large-Scale Immune Repertoire Data*. http://doi.org/10.5281/zenodo.3367200

Jahnke, K., Thiel, E., Martus, P., Herrlinger, U., Weller, M., Fischer, L., … on behalf of the German Primary Central Nervous System Lymphoma Study Group (G-PCNSL-SG). (2006). Relapse of primary central nervous system lymphoma: Clinical features, outcome and prognostic factors. *Journal of Neuro-Oncology*, *80*(2), 159–165. http://doi.org/10.1007/s11060-006-9165-6

Jang, I. K., Cronshaw, D. G., Xie, L.-k., Fang, G., Zhang, J., Oh, H., … Zou, Y. (2011). Growth-factor receptor-bound protein-2 (Grb2) signaling in B cells controls lymphoid follicle organization and germinal center reaction. *Proceedings of the National Academy of Sciences*, *108*(19), 7926–7931. http://doi.org/10.1073/pnas.1016451108

Jong, Mathilde Rikje Willemijn de, Langendonk, M., Reitsma, B., Herbers, P., Nijland, M., Huls, G., … Meerten, T. van. (2019). WEE1 Inhibition Enhances Anti-Apoptotic Dependency as a Result of Premature Mitotic Entry and DNA Damage. *Cancers*, *11*(11), 1743. http://doi.org/10.3390/cancers11111743

Jong, Mathilde R. W. de, Visser, L., Huls, G., Diepstra, A., Vugt, M. van, Ammatuna, E., … Meerten, T. van. (2018). Identification of relevant drugable targets in diffuse large B-cell lymphoma using a genome-wide unbiased CD20 guilt-by association approach. *PLOS ONE*, *13*(2), e0193098. http://doi.org/10.1371/journal.pone.0193098

Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., … Brown, J. R. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications*, *6*(1), 8866. http://doi.org/10.1038/ncomms9866

Kim, Sehui, Nam, S. J., Park, C., Kwon, D., Yim, J., Song, S. G., … Jeon, Y. K. (2019a). High tumoral PD-L1 expression and low PD-1 $^{+}$ or CD8 $^{+}$ tumor-infiltrating lymphocytes are predictive of a poor prognosis in primary diffuse large B-cell lymphoma of the central nervous system. *OncoImmunology*, *8*(9), e1626653. http://doi.org/10.1080/2162402X.2019.1626653

Kim, Sehui, Nam, S. J., Park, C., Kwon, D., Yim, J., Song, S. G., … Jeon, Y. K. (2019b). High tumoral PD-L1 expression and low PD-1 $^{+}$ or CD8 $^{+}$ tumor-infiltrating lymphocytes are predictive of a poor prognosis in primary diffuse large B-cell lymphoma of the central nervous system. *OncoImmunology*, *8*(9), e1626653. http://doi.org/10.1080/2162402X.2019.1626653

Kim, Sangtae, Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., … Saunders, C. T. (2018). Strelka2: Fast and accurate calling of germline and somatic variants. *Nature Methods*, *15*(8), 591–594.

http://doi.org/10.1038/s41592-018-0051-x

King, R. L., Goodlad, J. R., Calaminici, M., Dotlic, S., Montes-Moreno, S., Oschlies, I., ... Ferry, J. A. (2020). Lymphomas arising in immune-privileged sites: Insights into biology, diagnosis, and pathogenesis. *Virchows Archiv*, *476*(5), 647–665. http://doi.org/10.1007/s00428-019-02698-3

Kleinstern, G., Camp, N. J., Berndt, S. I., Birmann, B. M., Nieters, A., Bracci, P. M., ... Cerhan, J. R. (2020). Lipid Trait Variants and the Risk of Non-Hodgkin Lymphoma Subtypes: A Mendelian Randomization Study. *Cancer Epidemiology Biomarkers & Prevention*, *29*(5), 1074–1078. http://doi.org/10.1158/1055-9965.EPI-19-0803

Kleinstern, G., Yan, H., Hildebrandt, M. A. T., Vijai, J., Berndt, S. I., Ghesquières, H., ... Cerhan, J. R. (2020). Inherited variants at 3q13.33 and 3p24.1 are associated with risk of diffuse large B-cell lymphoma and implicate immune pathways. *Human Molecular Genetics*, *29*(1), 70–79. http://doi.org/10.1093/hmg/ddz228

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., ... Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. http://doi.org/10.1101/gr.129684.111

Kohli, R. M., Maul, R. W., Guminski, A. F., McClure, R. L., Gajula, K. S., Saribasak, H., ... Stivers, J. T. (2010). Local sequence targeting in the AID/APOBEC family differentially impacts retroviral restriction and antibody diversification. *The Journal of Biological Chemistry*, *285*(52), 40956–40964. http://doi.org/10.1074/jbc.M110.177402

Komohara, Y., Horlad, H., Ohnishi, K., Ohta, K., Makino, K., Hondo, H., ... Takeya, M. (2011). M2 Macrophage/Microglial Cells Induce Activation of Stat3 in Primary Central Nervous System Lymphoma. *Journal of Clinical and Experimental Hematopathology*, *51*(2), 93–99. http://doi.org/10.3960/jslrt.51.93

Komori, J., Marusawa, H., Machimoto, T., Endo, Y., Kinoshita, K., Kou, T., ... Chiba, T. (2008). Activation-induced cytidine deaminase links bile duct inflammation to human cholangiocarcinoma. *Hepatology (Baltimore, Md.)*, *47*(3), 888–896. http://doi.org/10.1002/hep.22125

Kondo, E., Ikeda, T., Izutsu, K., Chihara, D., Shimizu-Koresawa, R., Fujii, N., ... Suzuki, R. (2019). High-Dose Chemotherapy with Autologous Stem Cell Transplantation in Primary Central Nervous System Lymphoma: Data From the Japan Society for Hematopoietic Cell Transplantation Registry. *Biology of Blood and Marrow Transplantation*, *25*(5), 899–905. http://doi.org/10.1016/j.bbmt.2019.01.020

Kotlov, N., Bagaev, A., Revuelta, M. V., Phillip, J. M., Cacciapuoti, M. T., Antysheva, Z., ... Cerchietti, L. (2021). Clinical and Biological Subtypes of B-

cell Lymphoma Revealed by Microenvironmental Signatures. *Cancer Discovery*, *11*(6), 1468–1489. http://doi.org/10.1158/2159-8290.CD-20-0839

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, *27*(11), 1571–1572. http://doi.org/10.1093/bioinformatics/btr167

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., … Ma'ayan, A. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97. http://doi.org/10.1093/nar/gkw377

Kulis, M., Merkel, A., Heath, S., Queirós, A. C., Schuyler, R. P., Castellano, G., … Martín-Subero, J. I. (2015). Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nature Genetics*, *47*(7), 746–756. http://doi.org/10.1038/ng.3291

Kumar, V., Matsuo, K., Takahashi, A., Hosono, N., Tsunoda, T., Kamatani, N., … Matsuda, K. (2011). Common variants on 14q32 and 13q12 are associated with DLBCL susceptibility. *Journal of Human Genetics*, *56*(6), 436–439. http://doi.org/10.1038/jhg.2011.35

Kumari, N., Krishnani, N., Rawat, A., Agarwal, V., & Lal, P. (2009). Primary central nervous system lymphoma: Prognostication as per international extranodal lymphoma study group score and reactive CD3 collar. *Journal of Postgraduate Medicine*, *55*(4), 247. http://doi.org/10.4103/0022-3859.58926

Labreche, K., Daniau, M., Sud, A., Law, P. J., Royer-Perron, L., Holroyd, A., … Schmitt, A. (2019). A genome-wide association study identifies susceptibility loci for primary central nervous system lymphoma at 6p25.3 and 3p22.1: A LOC Network study. *Neuro-Oncology*, *21*(8), 1039–1048. http://doi.org/10.1093/neuonc/noz088

Langenbucher, A., Bowen, D., Sakhtemani, R., Bournique, E., Wise, J. F., Zou, L., … Lawrence, M. S. (2021). An extended APOBEC3A mutation signature in cancer. *Nature Communications*, *12*(1), 1602. http://doi.org/10.1038/s41467-021-21891-0

Langner-Lemercier, S., Houillier, C., Soussain, C., Ghesquières, H., Chinot, O., Taillandier, L., … Houot, R. (2016). Primary CNS lymphoma at first relapse/progression: Characteristics, management, and outcome of 256 patients from the French LOC network. *Neuro-Oncology*, *18*(9), 1297–1303. http://doi.org/10.1093/neuonc/now033

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., … Ding, L. (2012). SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, *28*(3), 311–317. http://doi.org/10.1093/bioinformatics/btr665

Lau, D., Bobe, A. M., & Khan, A. A. (2019). RNA Sequencing of the Tumor

Microenvironment in Precision Cancer Immunotherapy. *Trends in Cancer*, *5*(3), 149–156. http://doi.org/10.1016/j.trecan.2019.02.006

Law, P. J., Berndt, S. I., Speedy, H. E., Camp, N. J., Sava, G. P., Skibola, C. F., … Slager, S. (2017). Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nature Communications*, *8*(1), 14175. http://doi.org/10.1038/ncomms14175

Law, P. J., Sud, A., Mitchell, J. S., Henrion, M., Orlando, G., Lenive, O., … Houlston, R. S. (2017). Genome-wide association analysis of chronic lymphocytic leukaemia, Hodgkin lymphoma and multiple myeloma identifies pleiotropic risk loci. *Scientific Reports*, *7*(1), 41071. http://doi.org/10.1038/srep41071

Le Guyader-Peyrou, S. :, Defossez, G., Dantony, E., Mounier, M., Cornet, E., & Uhry, Z. (2019). *Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018* (Vol. Volume 2 – Hémopathies malignes). Saint-Maurice (Fra) : Santé publique France.

Le, M., Garcilazo, Y., Ibáñez-Juliá, M.-J., Younan, N., Royer-Perron, L., Benazra, M., … Alentorn, A. (2019). Pretreatment Hemoglobin as an Independent Prognostic Factor in Primary Central Nervous System Lymphomas. *The Oncologist*, *24*(9), e898–e904. http://doi.org/10.1634/theoncologist.2018-0629

Levy, O., DeAngelis, L. M., Filippa, D. A., Panageas, K. S., & Abrey, L. E. (2008). Bcl-6 predicts improved prognosis in primary central nervous system lymphoma. *Cancer*, *112*(1), 151–156. http://doi.org/10.1002/cncr.23149

Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., … Liu, X. S. (2016). Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biology*, *17*(1), 174. http://doi.org/10.1186/s13059-016-1028-7

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. http://doi.org/10.1093/bioinformatics/btp324

Li, Hanjie, Leun, A. M. van der, Yofe, I., Lubling, Y., Gelbard-Solodkin, D., Akkooi, A. C. J. van, … Amit, I. (2019). Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell*, *176*(4), 775–789.e18. http://doi.org/10.1016/j.cell.2018.11.043

Li, L., Su, N., Cui, M., Li, H., Zhang, Q., Yu, N., … Cao, Z. (2019). Activation-induced cytidine deaminase expression in colorectal cancer. *International Journal of Clinical and Experimental Pathology*, *12*(11), 4119–4124.

Li, Y., Yu, Z., Jiang, T., Shao, L., Liu, Y., Li, N., … Chang, Q. (2018). *Textless*span style="font-variant:small-caps;"\textgreaterSNCA\textless/span\textgreater , a novel biomarker for Group 4 medulloblastomas, can inhibit tumor invasion and induce apoptosis. *Cancer Science*, *109*(4), 1263–1275. http://doi.org/10.1111/cas.13515

Lin, E. Y., & Pollard, J. W. (2007). Tumor-Associated Macrophages Press

the Angiogenic Switch in Breast Cancer: Figure 1. *Cancer Research*, *67*(11), 5064–5066. http://doi.org/10.1158/0008-5472.CAN-07-0912

Linet, M. S., Vajdic, C. M., Morton, L. M., Roos, A. J. de, Skibola, C. F., Boffetta, P., … Chiu, B. C. H. (2014). Medical History, Lifestyle, Family History, and Occupational Risk Factors for Follicular Lymphoma: The InterLymph Non-Hodgkin Lymphoma Subtypes Project. *JNCI Monographs*, *2014*(48), 26–40. http://doi.org/10.1093/jncimonographs/lgu006

Linossi, E. M., & Nicholson, S. E. (2015). Kinase inhibition, competitive binding and proteasomal degradation: Resolving the molecular function of the suppressor of cytokine signaling (SOCS) proteins. *Immunological Reviews*, *266*(1), 123–133. http://doi.org/10.1111/imr.12305

Litchfield, K., Reading, J. L., Puttick, C., Thakkar, K., Abbosh, C., Bentham, R., … Swanton, C. (2021). Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell*, *184*(3), 596–614.e14. http://doi.org/10.1016/j.cell.2021.01.002

Liu, D., Schilling, B., Liu, D., Sucker, A., Livingstone, E., Jerby-Arnon, L., … Schadendorf, D. (2019). Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nature Medicine*, *25*(12), 1916–1927. http://doi.org/10.1038/s41591-019-0654-5

Liu, Y.-T., & Sun, Z.-J. (2021). Turning cold tumors into hot tumors by improving T-cell infiltration. *Theranostics*, *11*(11), 5365–5386. http://doi.org/10.7150/thno.58390

Liu, Y., Yao, Q., & Zhang, F. (2021). Diagnosis, prognosis and treatment of primary central nervous system lymphoma in the elderly population (Review). *International Journal of Oncology*, *58*(3), 371–387. http://doi.org/10.3892/ijo.2021.5180

Locke, W. J., Guanzon, D., Ma, C., Liew, Y. J., Duesing, K. R., Fung, K. Y. C., & Ross, J. P. (2019). DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Frontiers in Genetics*, *10*, 1150. http://doi.org/10.3389/fgene.2019.01150

Longato, E., Vettoretti, M., & Di Camillo, B. (2020). A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *Journal of Biomedical Informatics*, *108*, 103496. http://doi.org/10.1016/j.jbi.2020.103496

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. http://doi.org/10.1038/ng.2653

Lossos, I. S., Levy, R., & Alizadeh, A. A. (2004). AID is expressed in germinal center B-cell-like and activated B-cell-like diffuse large-cell lymphomas and is not correlated with intraclonal heterogeneity. *Leukemia*, *18*(11), 1775–1779. http://doi.org/10.1038/sj.leu.2403488

Louveau, A., Smirnov, I., Keyes, T. J., Eccles, J. D., Rouhani, S. J., Peske, J. D., ... Kipnis, J. (2015). Structural and functional features of central nervous system lymphatic vessels. *Nature*, *523*(7560), 337–341. http://doi.org/10.1038/nature14432

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. http://doi.org/10.1186/s13059-014-0550-8

Lu, X., Meng, J., Zhou, Y., Jiang, L., & Yan, F. (2021). *MOVICS* : An R package for multi-omics integration and visualization in cancer subtyping. *Bioinformatics*, *36*(22-23), 5539–5541. http://doi.org/10.1093/bioinformatics/btaa1018

Lv, C., Wang, J., Zhou, M., Xu, J.-Y., Chen, B., & Wan, Y. (2022). Primary central nervous system lymphoma in the United States, 1975–2017. *Therapeutic Advances in Hematology*, *13*, 204062072110661. http://doi.org/10.1177/20406207211066166

Machulla, H. K., Müller, L. P., Schaaf, A., Kujat, G., Schönermarck, U., & Langner, J. (2001). Association of chronic lymphocytic leukemia with specific alleles of the HLA-DR4:DR53:DQ8 haplotype in German patients. *International Journal of Cancer*, *92*(2), 203–207. http://doi.org/10.1002/1097-0215(200102)9999:9999<::aid-ijc1167>3.0.co;2-a

Makino, K., Nakamura, H., Hide, T., Kuroda, J., Yano, S., & Kuratsu, J. (2015). Prognostic impact of completion of initial high-dose methotrexate therapy on primary central nervous system lymphoma: A single institution experience. *International Journal of Clinical Oncology*, *20*(1), 29–34. http://doi.org/10.1007/s10147-014-0692-4

Makino, K., Nakamura, H., Kino, T., Takeshima, H., & Kuratsu, J.-I. (2006). Rising incidence of primary central nervous system lymphoma in Kumamoto, Japan. *Surgical Neurology*, *66*(5), 503–506. http://doi.org/10.1016/j.surneu.2006.05.055

Marcelis, L., Antoranz, A., Delsupehe, A.-M., Biesemans, P., Ferreiro, J. F., Debackere, K., ... Tousseyn, T. (2020). In-depth characterization of the tumor microenvironment in central nervous system lymphoma reveals implications for immune-checkpoint therapy. *Cancer Immunology, Immunotherapy*, *69*(9), 1751–1766. http://doi.org/10.1007/s00262-020-02575-y

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10. http://doi.org/10.14806/ej.17.1.200

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., ... Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, *171*(5), 1029–1041.e21. http://doi.org/10.1016/j.cell.2017.09.042

Mas-Ponte, D., & Supek, F. (2020). DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nature Genetics*, *52*(9), 958–968. http://doi.org/10.1038/s41588-020-0674-6

Matsumoto, Y., Marusawa, H., Kinoshita, K., Endo, Y., Kou, T., Morisawa, T., ... Chiba, T. (2007). Helicobacter pylori infection triggers aberrant expression of activation-induced cytidine deaminase in gastric epithelium. *Nature Medicine*, *13*(4), 470–476. http://doi.org/10.1038/nm1566

Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., ... Bolli, N. (2019). A practical guide for mutational signature analysis in hematological malignancies. *Nature Communications*, *10*(1), 2969. http://doi.org/10.1038/s41467-019-11037-8

Mayakonda, A., Schönung, M., Hey, J., Batra, R. N., Feuerstein-Akgoz, C., Köhler, K., ... Toth, R. (2021). Methrix: An R/Bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. *Bioinformatics*, *36*(22-23), 5524–5525. http://doi.org/10.1093/bioinformatics/btaa1048

McGranahan, Nicholas, Favero, F., Bruin, E. C. de, Birkbak, N. J., Szallasi, Z., & Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, *7*(283), 283ra54–283ra54. http://doi.org/10.1126/scitranslmed.aaa1408

McGranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., ... Swanton, C. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, *351*(6280), 1463–1469. http://doi.org/10.1126/science.aaf1490

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. http://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., ... Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. http://doi.org/10.1186/s13059-016-0974-4

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M. G. F., ... Shah, S. P. (2011). deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. *PLoS Computational Biology*, *7*(5), e1001138. http://doi.org/10.1371/journal.pcbi.1001138

Melsted, P., Hateley, S., Joseph, I. C., Pimentel, H., Bray, N., & Pachter, L. (2017). *Fusion detection and quantification by pseudoalignment* (preprint). Bioinformatics. Retrieved from http://biorxiv.org/lookup/doi/10.1101/166322

Mendez, J. S., Ostrom, Q. T., Gittleman, H., Kruchko, C., DeAngelis, L. M., Barnholtz-Sloan, J. S., & Grommes, C. (2018). The elderly left behind—changes in survival trends of primary central nervous system lymphoma over the past 4 decades. *Neuro-Oncology*, *20*(5), 687–694. http://doi.org/10.1093/neuonc/nox187

Menzel, U., Greiff, V., Khan, T. A., Haessler, U., Hellmann, I., Friedensohn, S.,

... Reddy, S. T. (2014). Comprehensive Evaluation and Optimization of Amplicon Library Preparation Methods for High-Throughput Antibody Sequencing. *PLoS ONE*, *9*(5), e96727. http://doi.org/10.1371/journal.pone.0096727

Methot, S. P., Litzler, L. C., Subramani, P. G., Eranki, A. K., Fifield, H., Patenaude, A.-M., ... Di Noia, J. M. (2018). A licensing step links AID to transcription elongation for mutagenesis in B cells. *Nature Communications*, *9*(1), 1248. http://doi.org/10.1038/s41467-018-03387-6

Meulen, M. van der, Dinmohamed, A. G., Visser, O., Doorduijn, J. K., & Bromberg, J. E. C. (2017). Improved survival in primary central nervous system lymphoma up to age 70 only: A population-based study on incidence, primary treatment and survival in the Netherlands, 1989–2015. *Leukemia*, *31*(8), 1822–1825. http://doi.org/10.1038/leu.2017.128

Miao, D., Margolis, C. A., Vokes, N. I., Liu, D., Taylor-Weiner, A., Wankowicz, S. M., ... Van Allen, E. M. (2018). Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nature Genetics*, *50*(9), 1271–1281. http://doi.org/10.1038/s41588-018-0200-2

Mills, M. C., & Rahal, C. (2019). A scientometric review of genome-wide association studies. *Communications Biology*, *2*(1), 9. http://doi.org/10.1038/s42003-018-0261-x

Milpied, P., Cervera-Marzal, I., Mollichella, M.-L., Tesson, B., Brisou, G., Traverse-Glehen, A., ... Nadel, B. (2018). Human germinal center transcriptional programs are de-synchronized in B cell lymphoma. *Nature Immunology*, *19*(9), 1013–1024. http://doi.org/10.1038/s41590-018-0181-4

Miyasato, Y., Takashima, Y., Takeya, H., Yano, H., Hayano, A., Nakagawa, T., ... Komohara, Y. (2018). The expression of PD-1 ligands and IDO1 by macrophage/microglia in primary central nervous system lymphoma. *Journal of Clinical and Experimental Hematopathology*, *58*(2), 95–101. http://doi.org/10.3960/jslrt.18001

Momeni, Z., Hassanzadeh, E., Saniee Abadeh, M., & Bellazzi, R. (2020). A survey on single and multi omics data mining methods in cancer data classification. *Journal of Biomedical Informatics*, *107*, 103466. http://doi.org/10.1016/j.jbi.2020.103466

Mondello, P., Cuzzocrea, S., Arrigo, C., Pitini, V., Mian, M., & Bertoni, F. (2020). STAT6 activation correlates with cerebrospinal fluid IL-4 and IL-10 and poor prognosis in primary central nervous system lymphoma. *Hematological Oncology*, *38*(1), 106–110. http://doi.org/10.1002/hon.2679

Monti, S. (2003). [No title found]. *Machine Learning*, *52*(1/2), 91–118. http://doi.org/10.1023/A:1023949509487

Moore, A., Kane, E., Wang, Z., Panagiotou, O. A., Teras, L. R., Monnereau, A., ... Berndt, S. I. (2020). Genetically Determined Height

and Risk of Non-hodgkin Lymphoma. *Frontiers in Oncology*, *9*, 1539. http://doi.org/10.3389/fonc.2019.01539

Morganella, S., Alexandrov, L. B., Glodzik, D., Zou, X., Davies, H., Staaf, J., ... Nik-Zainal, S. (2016). The topography of mutational processes in breast cancer genomes. *Nature Communications*, *7*(1), 11383. http://doi.org/10.1038/ncomms11383

Morin, A., Kwan, T., Ge, B., Letourneau, L., Ban, M., Tandre, K., ... Pastinen, T. (2016). Immunoseq: The identification of functionally relevant variants through targeted capture and sequencing of active regulatory regions in human immune cells. *BMC Medical Genomics*, *9*(1), 59. http://doi.org/10.1186/s12920-016-0220-7

Morton, L. M., Slager, S. L., Cerhan, J. R., Wang, S. S., Vajdic, C. M., Skibola, C. F., ... Sampson, J. N. (2014). Etiologic Heterogeneity Among Non-Hodgkin Lymphoma Subtypes: The InterLymph Non-Hodgkin Lymphoma Subtypes Project. *JNCI Monographs*, *2014*(48), 130–144. http://doi.org/10.1093/jncimonographs/lgu013

Mueller, L. P., & Machulla, H. K. G. (2002). Increased Frequency of Homozygosity for HLA Class II Loci in Female Patients with Chronic Lymphocytic Leukemia. *Leukemia & Lymphoma*, *43*(5), 1013–1019. http://doi.org/10.1080/10428190290021588

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., & Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*, *102*(5), 553–563. http://doi.org/10.1016/s0092-8674(00)00078-7

Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., & Bock, C. (2019). RnBeads 2.0: Comprehensive analysis of DNA methylation data. *Genome Biology*, *20*(1), 55. http://doi.org/10.1186/s13059-019-1664-9

Nayyar, N., White, M. D., Gill, C. M., Lastrapes, M., Bertalan, M., Kaplan, A., ... Batchelor, T. T. (2019). MYD88 L265P mutation and CDKN2A loss are early mutational events in primary central nervous system diffuse large B-cell lymphomas. *Blood Advances*, *3*(3), 375–383. http://doi.org/10.1182/bloodadvances.2018027672

Neefjes, J., Jongsma, M. L. M., Paul, P., & Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature Reviews Immunology*, *11*(12), 823–836. http://doi.org/10.1038/nri3084

Newman, Aaron M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., ... Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, *12*(5), 453–457. http://doi.org/10.1038/nmeth.3337

Newman, Aaron M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., ... Alizadeh, A. A. (2019). Determining cell type abundance and

expression from bulk tissues with digital cytometry. *Nature Biotechnology*, *37*(7), 773–782. http://doi.org/10.1038/s41587-019-0114-2

Nguyen, H., Shrestha, S., Draghici, S., & Nguyen, T. (2019). PINSPlus: A tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, *35*(16), 2843–2846. http://doi.org/10.1093/bioinformatics/bty1049

Nilsen, G., Liestøl, K., Van Loo, P., Moen Vollan, H. K., Eide, M. B., Rueda, O. M., … Lingjærde, O. C. (2012). Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics*, *13*(1), 591. http://doi.org/10.1186/1471-2164-13-591

Niparuck, P., Boonsakan, P., Sutthippingkiat, T., Pukiat, S., Chantrathammachart, P., Phusanti, S., … Atichartakarn, V. (2019). Treatment outcome and prognostic factors in PCNSL. *Diagnostic Pathology*, *14*(1), 56. http://doi.org/10.1186/s13000-019-0833-1

Nishiyama, A., & Nakanishi, M. (2021). Navigating the DNA methylation landscape of cancer. *Trends in Genetics*, *37*(11), 1012–1027. http://doi.org/10.1016/j.tig.2021.05

Nonaka, T., Toda, Y., Hiai, H., Uemura, M., Nakamura, M., Yamamoto, N., … Kinoshita, K. (2016). Involvement of activation-induced cytidine deaminase in skin cancer development. *The Journal of Clinical Investigation*, *126*(4), 1367–1382. http://doi.org/10.1172/JCI81522

Olszewski, A. J., & Castillo, J. J. (2013). Survival of patients with marginal zone lymphoma: Analysis of the Surveillance, Epidemiology, and End Results database. *Cancer*, *119*(3), 629–638. http://doi.org/10.1002/cncr.27773

Ostrom, Q. T., Gittleman, H., Liao, P., Vecchione-Koval, T., Wolinsky, Y., Kruchko, C., & Barnholtz-Sloan, J. S. (2017). CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. *Neuro-Oncology*, *19*(suppl_5), v1–v88. http://doi.org/10.1093/neuonc/nox158

Pardoll, D. M. (2012). The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer*, *12*(4), 252–264. http://doi.org/10.1038/nrc3239

Park, H. Y., Hong, Y.-C., Lee, K., & Koh, J. (2019). Vitamin D status and risk of non-Hodgkin lymphoma: An updated meta-analysis. *PLOS ONE*, *14*(4), e0216284. http://doi.org/10.1371/journal.pone.0216284

Pasqualucci, L., Bhagat, G., Jankovic, M., Compagno, M., Smith, P., Muramatsu, M., … Dalla-Favera, R. (2008). AID is required for germinal center–derived lymphomagenesis. *Nature Genetics*, *40*(1), 108–112. http://doi.org/10.1038/ng.2007.35

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE*, *7*(2), e30619. http://doi.org/10.1371/journal.pone.0030619

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017).

Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. http://doi.org/10.1038/nmeth.4197

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, *32*(5), 462–464. http://doi.org/10.1038/nbt.2862

PCAWG Evolution & Heterogeneity Working Group, PCAWG Consortium, Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., ... Van Loo, P. (2020). The evolutionary history of 2,658 cancers. *Nature*, *578*(7793), 122–128. http://doi.org/10.1038/s41586-019-1907-7

PCAWG Mutational Signatures Working Group, PCAWG Consortium, Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., ... Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. http://doi.org/10.1038/s41586-020-1943-3

Pender, A., Titmuss, E., Pleasance, E. D., Fan, K. Y., Pearson, H., Brown, S. D., ... Laskin, J. (2021). Genome and Transcriptome Biomarkers of Response to Immune Checkpoint Inhibitors in Advanced Solid Tumors. *Clinical Cancer Research*, *27*(1), 202–212. http://doi.org/10.1158/1078-0432.CCR-20-1163

Phelan, J. D., Young, R. M., Webster, D. E., Roulland, S., Wright, G. W., Kasbekar, M., ... Staudt, L. M. (2018). A multiprotein supercomplex controlling oncogenic signalling in lymphoma. *Nature*, *560*(7718), 387–391. http://doi.org/10.1038/s41586-018-0290-0

Phillips, T. J., Forero-Torres, A., Sher, T., Diefenbach, C. S., Johnston, P., Talpaz, M., ... Barr, P. M. (2018). Phase 1 study of the PI3Kδ inhibitor INCB040093 ± JAK1 inhibitor itacitinib in relapsed/refractory B-cell lymphoma. *Blood*, *132*(3), 293–306. http://doi.org/10.1182/blood-2017-10-812701

Pieper, K., Grimbacher, B., & Eibel, H. (2013). B-cell biology and development. *Journal of Allergy and Clinical Immunology*, *131*(4), 959–971. http://doi.org/10.1016/j.jaci.2013.01.046

Ponzoni, M., Berger, F., Chassagne-Clement, C., Tinguely, M., Jouvet, A., Ferreri, A. J. M., ... on Behalf of the International Extranodal Lymphoma Study Group (IELSG). (2007). Reactive perivascular T-cell infiltrate predicts survival in primary central nervous system B-cell lymphomas. *British Journal of Haematology*, *138*(3), 316–323. http://doi.org/10.1111/j.1365-2141.2007.06661.x

Preusser, M., Woehrer, A., Koperek, O., Rottenfusser, A., Dieckmann, K., Gatterbauer, B., ... Chott, A. (2010). Primary central nervous system lymphoma: A clinicopathological study of 75 cases. *Pathology*, *42*(6), 547–552. http://doi.org/10.3109/00313025.2010.508786

Qiao, Y., Zhou, Y., Wu, C., Zhai, K., Han, X., Chen, J., ... Lin, D. (2013). Risk of genome-wide association study-identified genetic variants for non-Hodgkin lymphoma in a Chinese population. *Carcinogenesis*, *34*(7), 1516–1519.

http://doi.org/10.1093/carcin/bgt082

Rahmatallah, Y., Emmert-Streib, F., & Glazko, G. (2016). Gene set analysis approaches for RNA-seq data: Performance evaluation and application guideline. *Briefings in Bioinformatics*, *17*(3), 393–407. http://doi.org/10.1093/bib/bbv069

Rennard, S. I. (2005). Challenges and Opportunities for Combination Therapy in Chronic Obstructive Pulmonary Disease. *Proceedings of the American Thoracic Society*, *2*(4), 391–393. http://doi.org/10.1513/pats.200504-046SR

Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., ... Chan, T. A. (2017). Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*, *171*(4), 934–949.e16. http://doi.org/10.1016/j.cell.2017.09.028

Ricard, D., Idbaih, A., Ducray, F., Lahutte, M., Hoang-Xuan, K., & Delattre, J.-Y. (2012). Primary brain tumours in adults. *The Lancet*, *379*(9830), 1984–1996. http://doi.org/10.1016/S0140-6736(11)61346-9

Robbiani, D. F., Deroubaix, S., Feldhahn, N., Oliveira, T. Y., Callen, E., Wang, Q., ... Nussenzweig, M. C. (2015). Plasmodium Infection Promotes Genomic Instability and AID-Dependent B Cell Lymphoma. *Cell*, *162*(4), 727–737. http://doi.org/10.1016/j.cell.2015.07.019

Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., ... Gordenin, D. A. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genetics*, *45*(9), 970–976. http://doi.org/10.1038/ng.2702

Robinson, J., Soormally, A. R., Hayhurst, J. D., & Marsh, S. G. E. (2016). The IPD-IMGT/HLA Database – New developments in reporting HLA variation. *Human Immunology*, *77*(3), 233–237. http://doi.org/10.1016/j.humimm.2016.01.020

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. http://doi.org/10.1093/bioinformatics/btp616

Roider, T., Seufert, J., Uvarovskii, A., Frauhammer, F., Bordas, M., Abedpour, N., ... Dietrich, S. (2020). Dissecting intratumour heterogeneity of nodal B-cell lymphomas at the transcriptional, genetic and drug-response levels. *Nature Cell Biology*, *22*(7), 896–906. http://doi.org/10.1038/s41556-020-0532-x

Rommel, P. C., Bosque, D., Gitlin, A. D., Croft, G. F., Heintz, N., Casellas, R., ... Robbiani, D. F. (2013). Fate Mapping for Activation-Induced Cytidine Deaminase (AID) Marks Non-Lymphoid Cells During Mouse Development. *PLoS ONE*, *8*(7), e69208. http://doi.org/10.1371/journal.pone.0069208

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., & Hacohen, N. (2015). Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*, *160*(1-2), 48–61. http://doi.org/10.1016/j.cell.2014.12.033

Rustad, E. H., Yellapantula, V., Leongamornlert, D., Bolli, N., Ledergor, G.,

Nadeu, F., ... Maura, F. (2020). Timing the initiation of multiple myeloma. *Nature Communications*, *11*(1), 1917. http://doi.org/10.1038/s41467-020-15740-9

Saito, Y., Koya, J., Araki, M., Kogure, Y., Shingaki, S., Tabata, M., ... Kataoka, K. (2020). Landscape and function of multiple mutations within individual oncogenes. *Nature*, *582*(7810), 95–99. http://doi.org/10.1038/s41586-020-2175-2

Saleh, R., Taha, R. Z., Toor, S. M., Sasidharan Nair, V., Murshed, K., Khawar, M., ... Elkord, E. (2020). Expression of immune checkpoints and T cell exhaustion markers in early and advanced stages of colorectal cancer. *Cancer Immunology, Immunotherapy*, *69*(10), 1989–1999. http://doi.org/10.1007/s00262-020-02593-w

Sampson, J. N., Wheeler, W. A., Yeager, M., Panagiotou, O., Wang, Z., Berndt, S. I., ... Chatterjee, N. (2015). Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. *Journal of the National Cancer Institute*, *107*(12), djv279. http://doi.org/10.1093/jnci/djv279

Samstein, R. M., Lee, C.-H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., ... Morris, L. G. T. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature Genetics*, *51*(2), 202–206. http://doi.org/10.1038/s41588-018-0312-8

Sanchez-Mazas, A. (2020). A review of HLA allele and SNP associations with highly prevalent infectious diseases in human populations. *Swiss Medical Weekly*. http://doi.org/10.4414/smw.2020.20214

Sapoznik, S., Bahar-Shany, K., Brand, H., Pinto, Y., Gabay, O., Glick-Saar, E., ... Levanon, K. (2016). Activation-Induced Cytidine Deaminase Links Ovulation-Induced Inflammation and Serous Carcinogenesis. *Neoplasia (New York, N.Y.)*, *18*(2), 90–99. http://doi.org/10.1016/j.neo.2015.12.003

Sasayama, T., Tanaka, K., Mizowaki, T., Nagashima, H., Nakamizo, S., Tanaka, H., ... Kohmura, E. (2016). Tumor-Associated Macrophages Associate with Cerebrospinal Fluid Interleukin-10 and Survival in Primary Central Nervous System Lymphoma (PCNSL): Tumor-Associated Macrophages in PCNSL. *Brain Pathology*, *26*(4), 479–487. http://doi.org/10.1111/bpa.12318

Saunders, P. M., Vivian, J. P., O'Connor, G. M., Sullivan, L. C., Pymm, P., Rossjohn, J., & Brooks, A. G. (2015). A bird's eye view of NK cell receptor interactions with their MHC class I ligands. *Immunological Reviews*, *267*(1), 148–166. http://doi.org/10.1111/imr.12319

Sava, G. P., Speedy, H. E., Di Bernardo, M. C., Dyer, M. J. S., Holroyd, A., Sunter, N. J., ... Houlston, R. S. (2015). Common variation at 12q24.13 (OAS3) influences chronic lymphocytic leukemia risk. *Leukemia*, *29*(3), 748–751. http://doi.org/10.1038/leu.2014.311

Sawai, Y., Kodama, Y., Shimizu, T., Ota, Y., Maruno, T., Eso, Y., ... Chiba, T. (2015). Activation-Induced Cytidine Deaminase Contributes to Pancreatic Tumorigenesis by Inducing Tumor-Related Gene Mutations. *Cancer Research*, *75*(16), 3292–3301. http://doi.org/10.1158/0008-5472.CAN-14-3028

Schaff, L. R., & Grommes, C. (2021). Update on Novel Therapeutics for Primary CNS Lymphoma. *Cancers*, *13*(21), 5372. http://doi.org/10.3390/cancers13215372

Schmidt, J., Smith, A. R., Magnin, M., Racle, J., Devlin, J. R., Bobisse, S., ... Gfeller, D. (2021). Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Reports Medicine*, *2*(2), 100194. http://doi.org/10.1016/j.xcrm.2021.100194

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. http://doi.org/10.1093/bioinformatics/b

Schmitz, R., Wright, G. W., Huang, D. W., Johnson, C. A., Phelan, J. D., Wang, J. Q., ... Staudt, L. M. (2018). Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *New England Journal of Medicine*, *378*(15), 1396–1407. http://doi.org/10.1056/NEJMoa1801445

Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., ... Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications*, *9*(1), 20. http://doi.org/10.1038/s41467-017-02391-6

Sehn, L. H., & Salles, G. (2021). Diffuse Large B-Cell Lymphoma. *New England Journal of Medicine*, *384*(9), 842–858. http://doi.org/10.1056/NEJMra2027612

Sewell, A. K. (2012). Why must T cells be cross-reactive? *Nature Reviews Immunology*, *12*(9), 669–677. http://doi.org/10.1038/nri3279

Shao, L., Xu, C., Wu, H., Jamal, M., Pan, S., Li, S., ... Wei, Y. (2021). Recent Progress on Primary Central Nervous System Lymphoma—From Bench to Bedside. *Frontiers in Oncology*, *11*, 689843. http://doi.org/10.3389/fonc.2021.689843

Sheffield, N. C., & Bock, C. (2016). LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics*, *32*(4), 587–589. http://doi.org/10.1093/bioinformatics/btv612

Shen, R., & Seshan, V. E. (2016). FACETS: Allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research*, *44*(16), e131–e131. http://doi.org/10.1093/nar/gkw520

Shiels, M. S., Pfeiffer, R. M., Besson, C., Clarke, C. A., Morton, L. M., Nogueira, L., ... Engels, E. A. (2016). Trends in primary central nervous system lymphoma incidence and survival in the U.S. *British Journal of Haematology*, *174*(3), 417–424. http://doi.org/10.1111/bjh.14073

Shimabukuro-Vornhagen, A., Draube, A., Liebig, T. M., Rothe, A., Kochanek, M., & Bergwelt-Baildon, M. S. von. (2012). The immunosuppressive factors IL-10, TGF-β, and VEGF do not affect the antigen-presenting function of CD40-

activated B cells. *Journal of Experimental & Clinical Cancer Research*, *31*(1), 47. http://doi.org/10.1186/1756-9966-31-47

Shimizu, T., Marusawa, H., Matsumoto, Y., Inuzuka, T., Ikeda, A., Fujii, Y., ... Chiba, T. (2014). Accumulation of somatic mutations in TP53 in gastric epithelium with Helicobacter pylori infection. *Gastroenterology*, *147*(2), 407–417.e3. http://doi.org/10.1053/j.gastro.2014.04.036

Shin, S.-H., Jung, K.-W., Ha, J., Lee, S. H., Won, Y.-J., & Yoo, H. (2015). Population-based Incidence and Survival for Primary Central Nervous System Lymphoma in Korea, 1999-2009. *Cancer Research and Treatment*, *47*(4), 569–574. http://doi.org/10.4143/crt.2014.085

Shinde, J., Bayard, Q., Imbeaud, S., Hirsch, T. Z., Liu, F., Renault, V., ... Letouzé, E. (2018). Palimpsest: An R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/bty388

Shugay, M., Bagaev, D. V., Turchaninova, M. A., Bolotin, D. A., Britanova, O. V., Putintseva, E. V., ... Chudakov, D. M. (2015). VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Computational Biology*, *11*(11), e1004503. http://doi.org/10.1371/journal.pcbi.1004503

Singh, R., Shaik, S., Negi, B., Rajguru, J., Patil, P., Parihar, A., & Sharma, U. (2020). Non-Hodgkin's lymphoma: A review. *Journal of Family Medicine and Primary Care*, *9*(4), 1834. http://doi.org/10.4103/jfmpc.jfmpc_1037_19

Skibola, C. F., Akers, N. K., Conde, L., Ladner, M., Hawbecker, S. K., Cohen, F., ... Bracci, P. M. (2012). Multi-locus HLA class I and II allele and haplotype associations with follicular lymphoma. *Tissue Antigens*, *79*(4), 279–286. http://doi.org/10.1111/j.1399-0039.2012.01845.x

Skibola, Christine F., Berndt, S. I., Vijai, J., Conde, L., Wang, Z., Yeager, M., ... Rothman, N. (2014). Genome-wide Association Study Identifies Five Susceptibility Loci for Follicular Lymphoma outside the HLA Region. *The American Journal of Human Genetics*, *95*(4), 462–471. http://doi.org/10.1016/j.ajhg.2014.09.004

Skibola, Christine F., Bracci, P. M., Halperin, E., Conde, L., Craig, D. W., Agana, L., ... Brown, K. M. (2009). Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nature Genetics*, *41*(8), 873–875. http://doi.org/10.1038/ng.419

Skibola, Christine F., Conde, L., Foo, J.-N., Riby, J., Humphreys, K., Sillé, F. C., ... Smedby, K. E. (2012). A meta-analysis of genome-wide association studies of follicular lymphoma. *BMC Genomics*, *13*(1), 516. http://doi.org/10.1186/1471-2164-13-516

Slager, S. L., Benavente, Y., Blair, A., Vermeulen, R., Cerhan, J. R., Costantini, A. S., ... Sanjose, S. de. (2014). Medical History,

Lifestyle, Family History, and Occupational Risk Factors for Chronic Lymphocytic Leukemia/Small Lymphocytic Lymphoma: The InterLymph Non-Hodgkin Lymphoma Subtypes Project. *JNCI Monographs*, *2014*(48), 41–51. http://doi.org/10.1093/jncimonographs/lgu001

Slager, Susan L., Rabe, K. G., Achenbach, S. J., Vachon, C. M., Goldin, L. R., Strom, S. S., ... Cerhan, J. R. (2011). Genome-wide association study identifies a novel susceptibility locus at 6p21.3 among familial CLL. *Blood*, *117*(6), 1911–1916. http://doi.org/10.1182/blood-2010-09-308205

Slager, Susan L., Skibola, C. F., Di Bernardo, M. C., Conde, L., Broderick, P., McDonnell, S. K., ... Houlston, R. S. (2012). Common variation at 6p21.31 (BAK1) influences the risk of chronic lymphocytic leukemia. *Blood*, *120*(4), 843–846. http://doi.org/10.1182/blood-2012-03-413591

Smedby, Karin E., Foo, J. N., Skibola, C. F., Darabi, H., Conde, L., Hjalgrim, H., ... Liu, J. (2011). GWAS of Follicular Lymphoma Reveals Allelic Heterogeneity at 6p21.32 and Suggests Shared Genetic Susceptibility with Diffuse Large B-cell Lymphoma. *PLoS Genetics*, *7*(4), e1001378. http://doi.org/10.1371/journal.pgen.1001378

Smedby, K. E., & Ponzoni, M. (2017). The aetiology of B-cell lymphoid malignancies with a focus on chronic inflammation and infections. *Journal of Internal Medicine*, *282*(5), 360–370. http://doi.org/10.1111/joim.12684

Speedy, H. E., Di Bernardo, M. C., Sava, G. P., Dyer, M. J. S., Holroyd, A., Wang, Y., ... Houlston, R. S. (2014). A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nature Genetics*, *46*(1), 56–60. http://doi.org/10.1038/ng.2843

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*(Dec), 583–617.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. http://doi.org/10.1073/pnas.0506580102

Swanton, C., McGranahan, N., Starrett, G. J., & Harris, R. S. (2015). APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. *Cancer Discovery*, *5*(7), 704–712. http://doi.org/10.1158/2159-8290.CD-15-0344

Swerdlow, Steven H., Campo, E., Harris, N. L., Jaffe, E. S., Pileri, S. A., Stein, H., ... others. (2008). *WHO classification of tumours of haematopoietic and lymphoid tissues* (Vol. 2). International agency for research on cancer Lyon, France.

Swerdlow, Steven H., Campo, E., Pileri, S. A., Harris, N. L., Stein, H.,

Siebert, R., ... Jaffe, E. S. (2016). The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*, *127*(20), 2375–2390. http://doi.org/10.1182/blood-2016-01-643569

Tabouret, E., Houillier, C., Martin-Duverneuil, N., Blonski, M., Soussain, C., Ghesquières, H., ... Hoang-Xuan, K. (2016). Patterns of response and relapse in primary CNS lymphomas after first-line chemotherapy: Imaging analysis of the ANOCEF-GOELAMS prospective randomized trial. *Neuro-Oncology*, now238. http://doi.org/10.1093/neuonc/now238

Takano, S., Hattori, K., Ishikawa, E., Narita, Y., Iwadate, Y., Yamaguchi, F., ... Matsumura, A. (2018). MyD88 Mutation in Elderly Predicts Poor Prognosis in Primary Central Nervous System Lymphoma: Multi-Institutional Analysis. *World Neurosurgery*, *112*, e69–e73. http://doi.org/10.1016/j.wneu.2017.12.028

Talevich, E., Shain, A. H., Botton, T., & Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Computational Biology*, *12*(4), e1004873. http://doi.org/10.1371/journal.pcbi.1004873

Teras, L. R., DeSantis, C. E., Cerhan, J. R., Morton, L. M., Jemal, A., & Flowers, C. R. (2016). 2016 US lymphoid malignancy statistics by World Health Organization subtypes: 2016 US Lymphoid Malignancy Statistics by World Health Organization Subtypes. *CA: A Cancer Journal for Clinicians*, *66*(6), 443–459. http://doi.org/10.3322/caac.21357

Terry M. Therneau, & Patricia M. Grambsch. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. http://doi.org/10.1111/1467-9868.00293

Tomkova, M., Tomek, J., Kriaucionis, S., & Schuster-Böckler, B. (2018). Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biology*, *19*(1), 129. http://doi.org/10.1186/s13059-018-1509-y

Touat, M., Li, Y. Y., Boynton, A. N., Spurr, L. F., Iorgulescu, J. B., Bohrson, C. L., ... Ligon, K. L. (2020). Mechanisms and therapeutic implications of hypermutation in gliomas. *Nature*, *580*(7804), 517–523. http://doi.org/10.1038/s41586-020-2209-9

Trowsdale, J., & Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annual Review of Genomics and Human Genetics*, *14*(1), 301–323. http://doi.org/10.1146/annurev-genom-091212-153455

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., ... Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419–1260419. http://doi.org/10.1126/science.1260419

Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., ... Brors, B. (2021). Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Research*, *31*(3), 448–460. http://doi.org/10.1101/gr.257246.119

Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., ... Garraway, L. A. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, *350*(6257), 207–211. http://doi.org/10.1126/science.aad0095

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, 11.10.1–11.10.33. http://doi.org/10.1002/0471250953.bi1110s43

Vasan, N., Razavi, P., Johnson, J. L., Shao, H., Shah, H., Antoine, A., ... Baselga, J. (2019). Double PIK3CA mutations in cis increase oncogenicity and sensitivity to PI3Kα inhibitors. *Science (New York, N.Y.)*, *366*(6466), 714–723. http://doi.org/10.1126/science.aaw9032

Vijai, J., Wang, Z., Berndt, S. I., Skibola, C. F., Slager, S. L., Sanjose, S. de, ... Nieters, A. (2015). A genome-wide association study of marginal zone lymphoma shows association to the HLA region. *Nature Communications*, *6*(1), 5751. http://doi.org/10.1038/ncomms6751

Vijai, J., Wang, Z., Berndt, S. I., Slager, S. L., Cerhan, J. R., Skibola, C., ... Nieters, A. (2014). Abstract 5071: A genome-wide association study suggests evidence of variants at 6p21.32 associated with marginal zone lymphoma. In *Epidemiology* (pp. 5071–5071). American Association for Cancer Research. http://doi.org/10.1158/1538-7445.AM2014-5071

Villa, D., Tan, K. L., Steidl, C., Ben-Neriah, S., Al Moosawi, M., Shenkier, T. N., ... Slack, G. W. (2019). Molecular features of a large cohort of primary central nervous system lymphoma using tissue microarray. *Blood Advances*, *3*(23), 3953–3961. http://doi.org/10.1182/bloodadvances.2019000989

Villano, J. L., Koshy, M., Shaikh, H., Dolecek, T. A., & McCarthy, B. J. (2011). Age, gender, and racial differences in incidence and survival in primary CNS lymphoma. *British Journal of Cancer*, *105*(9), 1414–1418. http://doi.org/10.1038/bjc.2011.357

Wang, M., & Xu, S. (2019). Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity*, *123*(3), 287–306. http://doi.org/10.1038/s41437-019-0205-3

Wang, Q., Kieffer-Kwon, K.-R., Oliveira, T. Y., Mayer, C. T., Yao, K., Pai, J., ... Robbiani, D. F. (2017). The cell cycle restricts activation-induced cytidine deaminase activity to early G1. *Journal of Experimental Medicine*, *214*(1), 49–58.

http://doi.org/10.1084/jem.20161649

Wang, S., Jia, M., He, Z., & Liu, X.-S. (2018). APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. *Oncogene*, *37*(29), 3924–3936. http://doi.org/10.1038/s41388-018-0245-9

Wang, S. S., Carrington, M., Berndt, S. I., Slager, S. L., Bracci, P. M., Voutsinas, J., … Skibola, C. F. (2018). HLA Class I and II Diversity Contributes to the Etiologic Heterogeneity of Non-Hodgkin Lymphoma Subtypes. *Cancer Research*, *78*(14), 4086–4096. http://doi.org/10.1158/0008-5472.CAN-17-2900

Wherry, E. J., & Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. *Nature Reviews Immunology*, *15*(8), 486–499. http://doi.org/10.1038/nri3862

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., … Nagarajan, N. (2012). LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, *40*(22), 11189–11201. http://doi.org/10.1093/nar/gks918

Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A., & Staudt, L. M. (2003). A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proceedings of the National Academy of Sciences*, *100*(17), 9991–9996. http://doi.org/10.1073/pnas.1732008100

Wright, G. W., Huang, D. W., Phelan, J. D., Coulibaly, Z. A., Roulland, S., Young, R. M., … Staudt, L. M. (2020). A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications. *Cancer Cell*, *37*(4), 551–568.e14. http://doi.org/10.1016/j.ccell.2020.03.015

Wu, Y.-C., Kipling, D., Leong, H. S., Martin, V., Ademokun, A. A., & Dunn-Walters, D. K. (2010). High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood*, *116*(7), 1070–1078. http://doi.org/10.1182/blood-2010-03-275859

Xia, Y., Rao, L., Yao, H., Wang, Z., Ning, P., & Chen, X. (2020). Engineering Macrophages for Cancer Immunotherapy and Drug Delivery. *Advanced Materials*, *32*(40), 2002054. http://doi.org/10.1002/adma.202002054

Xiong, B., Yang, Y., Fineis, F. R., & Wang, J.-P. (2019). DegNorm: Normalization of generalized transcript degradation improves accuracy in RNA-seq analysis. *Genome Biology*, *20*(1), 75. http://doi.org/10.1186/s13059-019-1682-7

Yamamoto, K., Venida, A., Yano, J., Biancur, D. E., Kakiuchi, M., Gupta, S., … Kimmelman, A. C. (2020). Autophagy promotes immune evasion of pancreatic cancer by degrading MHC-I. *Nature*, *581*(7806), 100–105. http://doi.org/10.1038/s41586-020-2229-5

Yang, X., Wang, Y., Sun, X., Bai, X., Cui, Q., Zhu, H., … Liu, Y. (2020).

STAT3 Activation Is Associated with Interleukin-10 Expression and Survival in Primary Central Nervous System Lymphoma. *World Neurosurgery*, *134*, e1077–e1084. http://doi.org/10.1016/j.wneu.2019.11.100

Yang, Z., Wong, A., Kuh, D., Paul, D. S., Rakyan, V. K., Leslie, R. D., … Teschendorff, A. E. (2016). Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biology*, *17*(1), 205. http://doi.org/10.1186/s13059-016-1064-3

Yin, W., Xia, X., Wu, M., Yang, H., Zhu, X., Sun, W., & Ge, M. (2019). The impact of BCL-2/MYC protein expression and gene abnormality on primary central nervous system diffuse large B-cell lymphoma. *International Journal of Clinical and Experimental Pathology*, *12*(6), 2215–2223.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., … Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, *356*(6337), eaaj2239. http://doi.org/10.1126/science.aaj2239

Yoshida, K., Shiraishi, Y., Chiba, K., Okuno, Y., Nakamoto-Matsubara, R., Koriyama, S., … Ogawa, S. (2016). Whole-Genome Sequencing of Primary Central Nervous System Lymphoma and Diffuse Large B-Cell Lymphoma. *Blood*, *128*(22), 4112–4112. http://doi.org/10.1182/blood.V128.22.4112.4112

Yu, C., Chen, L., Huang, C., Lin, V. C., Lu, T., Lee, C., … Bao, B. (2020). Genetic association analysis identifies a role for *ANO5* in prostate cancer progression. *Cancer Medicine*, *9*(7), 2372–2378. http://doi.org/10.1002/cam4.2909

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. http://doi.org/10.1089/omi.2011.0118

Yuan, X.-G., Huang, Y.-R., Yu, T., Xu, Y., Liang, Y., Zhang, X.-H., … Zhao, X.-Y. (2020). Primary central nervous system lymphoma in China: A single-center retrospective analysis of 167 cases. *Annals of Hematology*, *99*(1), 93–104. http://doi.org/10.1007/s00277-019-03821-9

Yuan, Y., Ding, T., Wang, S., Chen, H., Mao, Y., & Chen, T. (2021). Current and emerging therapies for primary central nervous system lymphoma. *Biomarker Research*, *9*(1), 32. http://doi.org/10.1186/s40364-021-00282-z

Yusufova, N., Kloetgen, A., Teater, M., Osunsade, A., Camarillo, J. M., Chin, C. R., … Melnick, A. M. (2021). Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. *Nature*, *589*(7841), 299–305. http://doi.org/10.1038/s41586-020-3017-y

Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, *33*(Web Server), W741–W748. http://doi.org/10.1093/nar/gki475

Zhang, Baihao, Vogelzang, A., Miyajima, M., Sugiura, Y., Wu, Y., Chamoto, K., … Fagarasan, S. (2021). B cell-derived GABA elicits IL-10+ macrophages to

limit anti-tumour immunity. *Nature*. http://doi.org/10.1038/s41586-021-04082-1

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. http://doi.org/10.1038/ncomms14049

Zhong, C., Cozen, W., Bolanos, R., Song, J., & Wang, S. S. (2019). The role of \textlessspan style="font-variant:small-caps;"\textgreaterHLA\textless/span\textgreater variation in lymphoma aetiology and survival. *Journal of Internal Medicine*, joim.12911. http://doi.org/10.1111/joim.12911

Zhou, Y.-H., Xia, K., & Wright, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, *27*(19), 2672–2678. http://doi.org/10.1093/bioinformatics/btr449

Zhou, Y., Liu, W., Xu, Z., Zhu, H., Xiao, D., Su, W., … Zhong, M. (2018). Analysis of Genomic Alteration in Primary Central Nervous System Lymphoma and the Expression of Some Related Genes. *Neoplasia*, *20*(10), 1059–1069. http://doi.org/10.1016/j.neo.2018.08.012

First Appendix

Supplementary Appendix

Methods

Patient samples

A total of 147 biopsies from treatment-naive fresh-frozen (FF; discovery cohort) and 93 formalin-fixed, paraffin-embedded (FFPE; validation cohort) tumor samples from immunocompetent Epstein-Barr negative PCNSL were recollected from different hospitals across France (see Table S1). All patients had a complete systemic evaluation to rule out secondary central nervous system (CNS) diffuse large B-cell lymphoma (DLBCL). Diagnoses were established at the reference institution by specialized pathologists. We obtained appropriate consent from relevant institutional review boards, which coordinated the consent process at each tissue-source site; written informed consent was obtained from all participants. The Pitié Salpêtirère Hospital ethics committee approved the study (Ile-de-France VI, N° DC-2009-957) and CNIL (DR-2013-279). All patients received high-dose methotrexate (HD-MTX) regimens according to French national "Lymphome oculo-cérébral, LOC" PCNSL network[1]. Moreover, 19/134 (14.2%; FF cohort) and 24/93 (25.8%; FFPE cohort) received intensive chemotherapy with autologous stem cell rescue (IC-ASCR).

Summary of clinical data results: The median age in both the FF cohort (68; 95% confidence interval [CI]=66-72) and the FFPE cohort (median=67; $CI_{95\%}$=63-70) is not significantly different (p=0.24; Wilcoxon-test). The male/female proportions in both cohorts resulted in the same 49/51%; as well as the median Karnofsky Performance Status (KPS) (70; $CI_{95\%}$=60-70). There was no significant difference between cohorts in either overall survival (OS; 19.9 months versus 22.4 months; p=0.274; Wilcoxon-test) or progression-free survival (PFS; 9.7 months versus 10.9 months; p=0.522; Wilcoxon-test). See Table S1.

Immunohistochemical analysis

FFPE tumor tissues from both FF and FFPE cohorts were available to perform immunohistochemistry on 4-µm thick sections. In brief, sections were deparaffinized and antigen retrieval was carried out by microwaving. Sections were then incubated with the working dilution of each antibody raised against the following proteins: CD20, CD10, BCL6, CD3, Ki67, and MUM1/IRF4 (see Table S1).

In situ hybridization for EBV

Epstein-Barr virus (EBV)-encoded small RNAs (EBERs) EBER-in situ hybridization was performed in a routine manner for detection of EBER1 and EBER2 small nuclear EBV-encoded RNA (800-2842, Ventana Medical Systems, Roche Diagnostics GmbH, Mannheim Germany), as previously described on FFPE tissue[2].

Multi-omic data integration for PCNSL molecular subtyping

We used six levels of omic-data (85 samples) for the identification of four PCNSL molecular subtypes with clinical implications including expression data from mRNA, somatic mutations, copy number aberration (CNA) events (per chromosomal arm), genes involved in fusion events, TCR/BCR clonotypes, and immune cell proportions (obtained by mRNA based immune deconvolution). Expression data were first transformed to variance stabilizing transformation (VST) counts representing 35,995 genes which were reduced to 2,087 genes by evaluating the univariate Cox regression effect on OS ($p < 0.05$). Immune cell fraction (values 0-1, obtained by Cibersortx)[3,4] was reduced from 22 variables to five by univariate Cox regression effect on OS. T-cell receptor (TCR) / B-cell receptor (BCR) clonotypes had a total of 166 variables, including the number of unique B/T clonotypes; *IGHV, IGHK, TRBV, TRAV* locus gene usage and different repertoire diversity estimator indexes, that were first z-scored and then reduced to 19 variables by univariate Cox regression. Average segments per chromosome arm were calculated using svpluscnv R package (v0.99.1)[5]. For the gene

mutation matrix, we considered mutations if classified as damaging according to PolyPhen/SIFT predictions[6–9] (‘‘probably_damaging’’/‘‘possibly_damaging’’ or ‘‘deleterious’’/‘‘deleterious_low_confidence’’, respectively); we additionally added *CDKN2A, HLA-A, and B2M* gene deletion status since it is important in lymphomas. mRNA expressed fusion partners were separated by the originating genes (independently of the direction) and a binary fusion matrix was constructed with 1 if the sample expressed a fusion in the gene or 0 otherwise. We reduced the gene mutation matrix from 6,061 variables to 31 by mutation frequency higher than 10%; meanwhile, the fusion matrix was reduced from 1,276 to 43 by filtering if present in at least 3 samples. The most appropriate cluster number was obtained by clustering prediction index (CPI) and Gaps-statistics analyses (Fig. S1). High robust clustering was obtained by consensus clustering (Fig. S2 from main article) resulting from 10 different multi-omics clustering algorithms (iClusterBayes, moCluster, CIMLR, IntNMF, ConsensusClustering, COCA, NEMO, PINSPlus, SNF, and LRA)[10–19]. Silhouette score[20] was calculated to measure sample similarity across the detected molecular subtypes (Fig. S3). Most of the above analyses are integrated in the R package “MOVICS”[21].

Summary of clustering results. Even though the CPI and Gaps-statistics analyses showed a higher average statistical value at 3 or 6 number of subtypes, the best consensus ensembles and average silhouette width results (Fig. S2-S3) were obtained using a multi-omic cluster number of 4.

Summary of survival results (FF multi-omic cohort). We identified four significant clusters (CS1 to CS4) that display different clinical outcomes in OS (Global log-rank $p < 0.001$, Figure 1B from main article). These observations remained significant after adjusting by age and KPS in Cox proportional hazard ratio multivariate models; for example, the cluster CS3 was independently associated with the worst prognosis (Hazard ratio [HR] = 17.98; $CI_{95\%}$ = 2.3-140.3; $p = 0.006$; Fig. S4A). Moreover, the four PCNSL subtypes also display a distinct global progression-free survival (PFS; global log-rank $p = 0.045$). However,

unlike OS, the only group presenting a clear difference was CS4 (median=42 months; $CI_{95\%}$=16.6-66.8) when compared against the other subtypes with univariate and multivariate models (Fig. S4B-C and Table S2).

Summary of clinical data results (FF multi-omic cohort). The tumor mutational burden (TMB) and the fraction of genome altered (CNA losses and gains) was significantly inferior only in CS2 compared to the others (median=1.72 mutations/Mb, p=0.03), but we did not observe significant differences in either the median number of predicted immunogenic neoantigens, fraction of canonical-AID (c-AID) mutations or number of patients receiving IC-ASCR (p=0.44, p=0.25, and p=0.32; Table S2 and Fig. S5).

Correction of paraffin fixation process degradation from FFPE RNA-seq data

Given that the use of sequencing data from large biorepositories of FFPE tissues can alleviate the limited availability of FF samples and also facilitate routine clinical diagnostics, we have performed supplemental analysis to determine whether 3' RNA-seq data obtained from FFPE PCNSL samples were comparable, to RNA-seq and WES data from paired FF samples from the same tumor (n=5), in terms of mapping quality, gene expression, and activated B-cell (ABC) or germinal center B-cell (GCB) molecular subtyping (see Cell-of-origin assignment section). We evaluated these terms using either the FFPE RNA-seq data directly or after a correction of the degradation due to the fixation process (using DegNorm software)[22].

Only 3' RNA-seq samples (FFPE) were subjected to trimming in which the quality and mapping of the reads were assessed before and after using the program FASTQC[23] and STAR (v2.7.2a)[24], respectively. Cutadapt (v1.18)[25] allowed trimming poly (A) sequences having a length from 10-19 (mismatch error of 10%) at the 3' site only, then we filtered sequences with less than 20 bases length to avoid alignment error during subsequent analysis. This process allowed us to visualize the data and was used to confirm the good quality of the reads before usage. RNA fastq files were aligned to the human reference

genome (GRCh38, release 97) using STAR (version 2.7.2a) for the remaining downstream analyses. For FF samples (pair-ended sequences), only the forward (R1) fastq files were used for alignment to reduce the batch effect. Mapping parameters were obtained from STAR outputs and compared for each FF-FFPE pair using Fisher's exact test. Median transcript integrity number (TIN) across all the transcripts, gene body coverage, and distribution of mismatches across reads were assessed with RSeQC[26]. TIN means between conditions were compared using a t-test. Then, read counts from FFPE data were divided into two sets: i) read counts obtained directly from aligning with STAR, and ii) corrected read counts obtained after adjusting degradation using DegNorm[22] (inputs are the resulting bams from STAR). Correlation plots within each FF-FFPE pair (FFPE corrected and not corrected for degradation) were done using $\log_2$ VST transformed reads. Principal Component Analysis (PCA) and covariance heatmap were performed using the 'plotPCA' function of the DESeq2 package (default parameters)[27] or using ComplexHeatmap[28] (default Euclidean distance).

Summary of results. First, we looked at the alignment efficiency between 3' FFPE RNA-seq data and the FF RNA-seq data. We observed that the percentage of correctly mapped reads was not significantly different in the samples (except for sample 7833, Fig. S6A). Next, we evaluated the RNA-seq reads coverage over the gene body (Fig. S6B) and found all FFPE samples having higher degradation (lower coverage) towards the 5' region and up to the 75 percentile; on the other hand, FFPE samples had low coverage up to the 15 percentile. Furthermore, when we looked at the mismatch profiles (Fig. S6C), we observed a clear higher number of G>A and C>T transitions for the FFPE samples, indicative of chemical artifacts caused during the paraffin fixation process. Furthermore, sample 7833_FFPE shows a completely different pattern indicating, along with the found low unique mapping reads and total reads, that this sample was highly degraded.

Having found that the fixation process produces a high number of artifacts, we sought to apply an RNA-seq normalization pipeline to correct transcript degradation bias. To do so,

first, we performed Spearman correlation analyses using the VST gene read counts between the FF-FFPE pairs with either the uncorrected FFPE data or the corrected FFPE data. The not corrected FFPE data showed a median Spearman Rho of 0.66 while the corrected data had a median of 0.84 (p-value < 0.05, Figure S7); furthermore, as expected the sample 7833 had the lowest correlation value. Then we evaluated if the corrected/uncorrected data could be used to correctly assign the Cell-of-origin (COO, ABC or GCB subtypes) of each tumor and found that all samples with both data types can be correctly assigned (Figure S8A). Finally, by using PCA analysis we found that FF and FFPE samples clustered together when applied to uncorrected FFPE data. Interestingly, when correcting for degradation FF-FFPE pairs are clustered together (Figure S8B-C). In summary, we showed that DegNorm can correct most of the degradation signal from the FFPE process in PCNSL samples and that mismatch profiles can help discriminate against highly degraded samples.

Using transcriptomic data for assigning multi-omic defined PCNSL subtypes in FF and FFPE samples

Top 100 unique upregulated biomarkers (no overlapping across subtypes, Table S3) were identified for each molecular subtype (Deseq2 package v1.32.0) with a threshold of adjusted p-value < 0.05. For extrapolating molecular classification using one level omic-data, on one hand, these biomarkers were further used to apply the nearest template prediction (NTP)[29] method on mRNA expression data coming from a validation FFPE cohort (n = 93, data previously corrected for paraffin fixation process degradation); on the other hand, the partition around medoids (PAM) method was applied on mRNA coming from the complete set of FF-RNA (n = 123)[29].

Summary of results. A Cohen's kappa coefficient of 0.90 was obtained (p<0.001) when evaluating the accuracy of correctly assigning patients from the multi-omic cohort (Fig. S9). Next, we assigned the corresponding cluster subtype to the complete set of FF-RNA (n=123) finding 55 CS1, 26 CS2, 18 CS3 and 24 CS4 PCNSL subtypes which showed the same

clinical outcome behaviors. Specifically, they exhibited a global log-rank p-value of 0.002 (Fig. S10A) from which patients in the CS4 had the longest OS (median=24.3 months; $CI_{95\%}$=16.5-59.9) and those in the CS3 had the shortest (median=13.8 months; $CI_{95\%}$=5.1-16.2). Such observations remained significant after adjusting by age and KPS in Cox proportional hazard ratio multivariate models (Fig. S10B). Assigning a molecular subtypes using FFPE-RNA data was also possible in a cohort of 93 samples (Fig. S11), we found 21 CS1 and CS2, 16 CS3, and 35 CS4 PCNSL subtypes. Again, CS4 patients showed to have the longest survival, followed by CS1, CS2 and finally CS3 at both univariate and multivariate models (Figure 1C from main article and Fig. S12A). Additionally, in this cohort the CS4 subtype was also independently associated with a better response when considering PFS in univariate and multivariate models (Fig. S12B-C).

Whole exome sequencing (WES)

DNA extraction and Exome sequencing

DNA was extracted using Blood & Cell culture DNA mini kit (Qiagen) following the manufacturer's instructions. WES libraries were generated using the KaPa hyperprep kit in combination with the HyperExome capture kit (Roche). Libraries were sequenced on an illumina Nova-seq 6000 sequencing system using 2x100 bp paired-end sequencing (43 Mb couverture).

Alignment and quality control

Preprocessing of read alignments for analysis was done using Broad Institute's data processing pipeline with Picard's tools [30]. Fastq data were aligned using BWA-MEM[31] (version 0.7.17) with the same reference genome (GRCh38). The resulting aligned reads were processed to add read groups, sort, mark duplicates, create an index, realign around known indels, reassign mapping qualities, and recalibrate base quality scores [32–34].

Mutation calling

Variant calling was done using five different software (MuTect2 v4.1.9.0, Strelka2 v2.9.10, VarScan2 v2.4.4, Lofreq v2.1.3.1, and SomaticSniper v1.0.5.0)[32,35–38], then for common variants between all software were picked using SomaticSeq (v3.5.1)[39], and finally the indel and SNV vcfs were merged with GATK[32]. 98 samples were tumor-normal pairs from which a panel of normals (PON, GATK function from version 3.8.1.0)[32] was constructed to filter germline variants from this set or the Tumor-only samples (n=17). Mutations were annotated using Variant Effect Predictor (VEP, v98.2)[40], then transformed into a mutation annotation format (MAF) file and annotated with CNA data using facetsSuite (v2.0.8)[40,41]. The annotated MAF was further analyzed with maftools[42].

Summary of results. We identified 32,544 mutations in the 115 PCNSL samples analyzed of which 80.6% were nonsynonymous exonic variants (median=3.23 mutations/Mb; range = 0.02-85.49). Samples exhibited more deletions (median=1,045; range=0-4,863) than amplifications (median=285; range=0-959; (Table S4 and Fig. S13).

Attributing mutations as driver genes

Positively-selected genes were obtained by calculating dN/dS likelihood ratios (*dNdScv* package)[43] that quantifies the mode and strength of selection by comparing synonymous substitution rates (dS)—assumed to be neutral—with nonsynonymous substitution rates (dN). This analysis is done through a negative binomial regression modeling of the background mutation rate of each gene using distinct genomic covariates including variation in mutation density across genes, context-dependent substitutions (mutational signatures), transcriptional strand bias, chromatin state, expression and replication time. Additionally, we removed ultra-hypermutator samples and extremely mutated genes per sample, to avoid loss of sensitivity. Genes were considered as drivers (n = 466) if having q-values < 0.001 (Benjamini-Hochberg's multiple testing correction of p-values) [43].

Summary of results. We identified driver mutations, on the one hand, the hallmarks

mutations of PCNSL like *MYD88* (64%), *PIM1*(59%), *PRDM1* (57%), *GRHPR* (50%), *HLA-A/B/C* (49%, 30% and 13%), *BTG2* (47%), *CD79B* (43%), *TOX* (39%), *OSBPL10* (33%), *IRF4* (30%), *CDKN2A* (28%), *CD58* (27%), *ETV6* (26%), *MPEG1* (26%), *TBL1XR1* (25%), *KLHL14* (25%), *BTG1* (23%), *KMT2D* (23%), *CARD11* (22%), and *HIST1H1E* (18%) which are involved in BCR-TLR mediated NF-kB signaling, antigen presentation, cell-cycle, histone modification and B-cell differentiation regulation (Fig. S14)[44–46].

Tracking AICDA-related mutations

We used a code to detect AID-related mutations over *wrCy/rGyw* (+/- strand, where "W" stands to either adenine or thymine, "R" to purine, and "Y" to pyrimidine) motifs, giving a total of 8 motifs per strand (positive-strand = AACC, AACT, AGCC, AGCT, TACC, TACT, TGCC, TGCT; negative-strand = TTGG, TTGA, TCGG, TCGA, ATGG, ATGA, ACGG, ACGA), and its enrichment around a 60 bp flanking sequences (the code allows other bp windows for the user). Mutations were tagged as AICDA or not AICDA if overlapping or not with the mutations found in the output MAF after applying the function, as previously described [47].

Summary of results. We detected c-AID off-target mutations and found they represent 7.9% (6.8-8.5% at 95% CI) of SNV mutations which is higher than those provoked by APOBEC ($p < 0.001$, Wilcoxon-test, Fig. S15). Regarding their distribution, they fall within driver genes like *PIM1* (47%), *OSBPL10* (22%), *BTG1* (17%), *KLHL14* (15%), *CD79B* (10%), *IRF4* (9%), and *HIST1H1E* (6%) (Table S6 and Fig. S16). Furthermore, mutations in *PIM1* and *KLHL14* were found to be consecutive with an average mutational distance ≤1 Kb in some samples (named as Kataegic target in Fig. S16).

Attributing mutations to mutagenic processes

Mutations previously tagged as not AID were subjected to signature attribution to 46 (we excluded the single base substitution (SBS) signatures: 27, 39, 43, 45-60 since they are attributed to sequencing artifacts) of the 65 COSMIC mutational signatures (v3.0) using

Palimpsest package with default parameters [48–50]. Signatures were initially decomposed using non-negative matrix factorization with the Brunet algorithm[51] to extract the optimal number of signatures and then compared to the COSMIC signatures. To avoid over-fitting, the resulting signatures not showing a cosine score > 0.6 were removed and mutations were re-fitted using the remaining signatures[49]. Furthermore, signatures proportions per sample were re-calculated adding the number of previously identified AID mutations to the signature data. For double base substitutions (DBS) and indels (ID), signatures not contributing with at least one mutation within 50% of the samples per tumor type (Median-Signature$_n$ < 1) were removed and mutations were re-fitted using the remaining signatures. CNV signatures were derived using the sigminer R package (v2.0.4)[52] where the number of signatures was determined using the indication of stability from the cophenetic correlation coefficient plot (performing 30 runs); furthermore, samples were clustered into four different groups based on the consensus matrix resulting from multiple NMF runs where each group is specified by the most enriched copy number signature derived previously[52]. The attribution was run separately for hypermutated samples (n = 18).

Kataegic events

Kataegic events were identified using the KataegisPortal (v1.0.3)[53] and defined as having four or more consecutive mutations with an average mutational distance ≤1 Kb, excluding immune hypermutated regions[53]. Moreover, if the event was produced by only AID mutations it was classified as an AID kataegic event.

Timing of mutations and copy number gains

Timing categorization of mutations and copy number gains was done only for samples from which allele-specific integer copy numbers, ploidy, and purity estimates were available (n = 115). We used MutationTimeR (v1.00.2; n bootstraps = 200) as previously described[54], in brief for each mutation, assuming read counts follow a beta-binomial distribution, the mutation copy number (m) along with the clonal frequency (f) was determined inputting the

variant allele fraction, tumor purity, minor allele copy numbers (c), total copy numbers (C), sex and tumor heterogeneity (mutations clusters determined by FACETS v0.5.14)[55], then a mutation was assigned as subclonal (f < 1), early clonal (f = 1; m > 1), late clonal (f = 1; m = 1; C > 2: c = 0) or clonal (unspecified) otherwise. Timing of copy number loss of heterozygosity (LOH), mono-allelic, and bi-allelic gains was inferred using the number of mutations occurring at each allelic copy number[54]. Next, we used this calculated clonal allelic status for each mutation along with the mutational process that probably originated them to gain information about the relative timing of these processes. We performed a Wilcoxon matched-pairs signed rank test for paired data (per sample) to compare the proportion of mutations happening in clonal vs subclonal pairs, or early clonal vs late clonal pairs.

Summary of results. We observed that defective homologous recombination-based DNA damage repair (Cosmic signature SBS3), SBS5, and non c-AID (SBS9) are significantly more active at early clonal stages (p = 0.002 and 0.018, respectively), meanwhile, c-AID and cell division related signature (SBS1) are significantly more active at clonal stages (p = 0.007 and < 0.001, respectively) (Fig. S17-18).

Neoepitope analysis

Variant call format (VCF) files (n=91) were prepared according to pVAC-Seq's recommendations: 1) annotate them using Variant Effect Predictor (VEP, v98.2) with the Down-stream, Wildtype VEP plugins and transcript version parameter; 2) add expression data using the vatools' 'vcf-expression-annotator' [56,57]**.** pVAC-Seq (v4.0.8) was run for each patient and allele combination using tumor and normal protein epitopes of 8, 9, 10, and 11 amino acids in length (produced by reconstructing the nucleotide sequence surrounding the mutation using its coordinates from the VCF file) and using the human major histocompability complex (MHC) binding predictions generated by the Immune Epitope Database and Analysis Resource (IEDB, version 3.10.0) MHC class I and II prediction tools, MHCflurry and MHCnuggets (I and II)[58–60]. The median of all prediction values was used for

alleles with multiple prediction methods. We further filtered using a cutoff for the mutant IC50 binding score of 500 nmol/L. To account for immunogenicity based on the prediction of neopeptide TCR recognition and neopeptide HLA binding, PRIME software was run with default parameters. Neoepitopes were classified as “Immunogenic” if having a PRIME %rank score (the fraction of random 700,000 8- to 14-mers that would have a score higher than the peptide provided in input) lower or equal to 0.5% for the corresponding HLA haplotype of the patient where the neopeptide occurred, or as “Non-Immunogenic” otherwise [61].

Copy number alterations calling

We calculated absolute copy number calling with segmentation values, tumor purity and ploidy using FACETS (v0.5.14, input GATK common SNP locations: ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/GATK/00-common_all.vcf.gz) for paired samples where heterozygous sites were called from the paired normal, while for the tumor-only samples the log-ratio noise was minimized using a pool panel of normals (n = 98) through CNVkit (v0.9.9) [55,62]. Chromosome instability index was calculated as the fold change (∓0.2 segmean) CNV log-ratio between each sample and the reference using svpluscnv R package[5].

Significance analysis of recurrent SCNAs using GISTIC2.0

Arm-level and focal peaks of recurrent copy number alterations were identified from the results of Allelic Capseg using GISTIC2.0 (version 23) as previously described[63]. Regions with germline copy number variants were excluded from the analysis. Events with a q-value of less than 0.1 were reported as significant. We specified a 95% confidence interval to determine wide peak boundaries. Samples were run separately if belonging to the low or high mutated group, for a total of 108 samples with available information.

Summary of results. We identified significant recurrent amplifications in 12q15 (53%), 11p15.4 (40%), 18q21.33 (42%), 19p13.13 (34%), 11q23.3 (27%), and deletions in

14q32.33 (84%), 2p11.2 (55%), 6p21 (39%), 6q21 (65%), 6q27 (49%), and 9p21.3 (28%). Such events affected genes like *CD58, IL10, IL20, IL24, BTG2, IRF4, CD83, CD24, PRDM1, MAP3K5, HLA-A/B/C, CARD11, IL6, TOX, CDKN2A, CXCR5, ETV6, CD3G, B2M,* among others (Fig. S19).

PCNSL recurrent alterations comparisons with systemic DLBCL

We directly compared the genetic landscape of our cohort to the previously published DLBCL data[46]. Using mirror bar plots (Figure 2D from main article), the frequencies of the recurrent genetic alterations in PCNSL were compared to those in DLBCL[46]. An asterisk denotes the known genetic driver events in DLBCL (Figure 2D from main article). Additionally, we also compared the recurrent CNAs in the PCNSL and DLBCL cohorts using GISTIC[63] mirror plots (Figure 2B from main article).

Summary of results. PCNSL and DLBCL samples exhibited significant arm 6q deletions along with amplifications in arms 1q, 7p, 7q, 11q, 12q, 18q, 19p, and 21q. We found additional, not previously described, amplifications in 1q32.1 (33%, *IL10*), 11q23.3 (26%, *CD3G*), and deletions in 6p25.3 (21%, *IRF4*), 22q11.22 (29%, *GGTLC2*) and 14q32.33 (84%; Fig. S20)

Integrative analysis of gene expression and copy number data

Samples with RNA-seq normalized gene counts (VST method, DESeq2 package) and SCNA GISTIC data were used (n=80) to infer gene expression alterations due to CNA events. For each of the 66 GISTIC-defined alteration peaks (false discovery ratio [FDR] q-value <0.1), samples were grouped as “affected” or “not affected” and genes were subset to the ones altered by the evaluated peak, then a differential expression analysis using R package DESeq2 was launched (FDR correction for p-values) [27]. We filtered out genes having an opposite association since we were interested in gene expression up-regulation among samples with copy gain and gene expression down-regulation among samples with copy

loss. Genes with FDR<0.1 and absolute fold change >1.0 were considered significantly affected. Representation of this information along with mutated genes in regions of SCNA was done using a modified gisticChromPlot function from maftools R package (Fig. S19 and S21)[42].

Summary of results. Genes affected by either gains or losses include *CD3G* (FC= 1.25), *IRF4* (FC=-1.03), *CCL21* (FC= -1.39), and *GGTLC2* (FC=-1.76, FDR q-value<0.1). The complete list of genes is given in Table S7 and a representation in Fig. S19 and S21.

Distinct genetic signatures within PCNSL subtypes

Marker genes associated with each molecular cluster were identified using a Fisher test (one-sided) where we evaluated the presence or absence of an event (mutation, CNA, or fusion) within-cluster vs outside-cluster and corrected the observed p-values by FDR (q-value ≤0.1 selected, Figure 2C from main article). SBS, DBS, and ID signature activity were tested across groups (Wilcoxon-test).

Summary of results. We only found the mutational processes DBS4 and ID8 to be more active in clusters CS4 and CS3, respectively (Wilcoxon-test; Fig. S22).

RNA-seq

RNA-extraction and sequencing

Total RNA from FF samples (n = 123) were extracted using trizol and Guanidine Isothiocyanate and quality was assessed using a NanoDrop spectrophotometer (Thermo Fisher Scientific) and electropherogram profiles on an Agilent Bioanalyzer (Agilent Technologies), as previously described[64]**.** For the validation FFPE cohort (n = 93) tissue sections were deparaffinized and processed with the RSC RNA FFPE Kit (Maxwell cat # AS1440) to recover a final elution of 50 µL in nuclease-free water. RNA quantity and quality was measured via Agilent 2200 TapeStation R6K ScreenTape assay (cat # 5067–5367).

RNA integrity, determined by the RNA integrity number (RIN), was determined with the 2100 Bioanalyzer (Agilent).

mRNA (20 ng) from FF samples was used for whole-transcriptome libraries preparation with the KAPA mRNA HyperPrep Kits (KAPA Biosystems/Roche). Equal concentrations of cDNA library from each sample were pooled for sequencing on an illumina Nova-seq 6000 sequencing system using 2x150 bp paired-end sequencing.

For the validation FFPE cohort 3' RNA-seq was performed where libraries were constructed using QuantSeq 3' mRNA-Seq Library Prep Kit (Lexogen, Austria) according to the manufacturer's instructions. The pooled 3′mRNA-Seq libraries were loaded to an illumina Nova-seq 6000 sequencing system using single-end 100 bp sequencing.

Alignment and quantification

RNA fastq files were aligned to the human reference genome (GRCh38, release 97) using STAR (version 2.7.2a) for mapping quality assessment after trimming of adaptor sequences (Cutadapt v1.18) [24,25]. Gene read counts were corrected for paraffin fixation process degradation using DegNorm, only for the FFPE validation cohort[22], or used directly as input for differential expression analysis using DESeq2 (FDR p-value <0.05, absolute fold change >1 and average read count >4). Counts were transformed using the variance-stabilizing transformation (VST) function in DESeq2 for downstream analyses[27].

Cell-of-origin (COO) assignment

Two sets of previously described gene panels were used to classify samples into ABC (*SH3BP5*, *IRF4*, *PIM1*, *ENTPD1*, *BLNK*, *CCND2*, *ETV6*, *FUT8*, *BMF*, *IL16*, *PTPN1*) or GCB (*ITPKB*, *MMEBCL6*, *MYBL1*, *DENND3*, *NEK6*, *LMO2*, *LRMP*, *SERPINA9*) subgroups[65]. Gene expression, from VST transformed reads, was quantile normalized, log2 transformed, and then z-normalized for all genes. Then ABC/GCB genes were separated and scores were computed for each sample by taking the average of the z-scores for each gene panel. The difference in the ABC-specific score to the GCB-specific score is computed to obtain the

RNAseq subtype score which is used to classify the samples as ABC if the RNAseq subtype score is >0.25 and its GCB score is <0.75, and it is classified as GCB if the RNAseq subtype score is < -0.25 and its ABC score<0.75. Otherwise, the samples belong to an unclassified group[66].

Summary of results. From the 123 FF samples 108 were ABC subtype, 5 GCB, and 10 unclassified; on the other hand, from the 93 FFPE samples 80 were ABC, 7 GCB, and 6 unclassified.

HLA haplotyping

Each sample, three-digit class I and II HLA type, was determined from FASTQ files using arcasHLA (default parameters, version 3.24.0, assembly GRCh38)[67].

Pathway analysis

Gene set enrichment analysis (GSEA)

GSEA analysis was performed using the GO/Reactome/Hallmark database through R package clusterProfiler [68] and applying Bonferroni correction (q-value <0.05).

Gene set variation analysis (GSVA)

GSVA[69] was applied through MOVICS using a list 23 of tumor microenvironment (TME) gene set list previously described for lymphoma (Fges) [70]; the cytolytic activity was done using the expression of *GZMB* and *PRF1*[71]; PCR2 complex activity was scored using *EED, PCGF1, EZH2, PCL, SUZ12, PCL,* and *JARID2*[72,73]; glycolytic activity, MYC induction, B-cell differentiation, and transcription factors activity gene signatures were described by Wright et al. and Milpied et al.[74,75]. Biological oncogenic pathways were calculated using activity scores from PROGENy package (v1.14.0)[76]. These signatures were used, in both the FF multi-omic cohort (n = 85) and the FFPE cohort (n = 93), to determine whether or not the observed distinct mutations within each molecular subtype could affect oncogenic pathways and B-cell

differentiation programs transcriptionally. Scores were compared by Wilcoxon test where p-values were adjusted for multiple testing by FDR.

Summary of results. From the Fges signatures we observed the CS1 cluster with higher B-cell content, cell proliferation, and PRC2 complex activity while the CS2 cluster exhibited higher gamma-Aminobutyric acid (GABA) synthesis and expression of vascular endothelial cells (VEC). The CS4 group showed higher expression of signatures related to inflammation such as immune checkpoints, MHC-I, MHC-II, cytolytic activity, macrophages M1, NK cells, fibroblastic reticular cells (FRC), and follicular dendritic cells (FDC), and tumor infiltrating lymphocytes (TILs) (Fig. S23). Regarding oncogenic (Fig. S24) and PROGENy's (Pathway RespOnsive GENes, Fig. S25) signatures, the CS1 group showed higher MYC induction, glycolytic and PI3K activity; the CS2 group was characterized by higher p53 activity; and the CS3/CS4 groups exhibited higher MAPK, NF-kB, and JAK-STAT activities. Concerning the B-cell differentiation programs, the CS1 group expressed a mixture of GC cells (ligh, intermediate and dark zone) which is consistent with the observed MYC induction activity but not with the expression of transcription factors (TFs) that upregulates *IRF4*, *TCF3* and downregulates *BCL6*. However, these genes are preferentially disturbed in CS1 by 6p25.3, 19q13.12 deletions, and *BCL6* mutations (Figure 2D from main article). The cluster CS4 presents an enrichment in terminally differentiated plasma cells that goes in line with upregulated BCL6 repression and the absence of *MYC* induction and *BCL6* mutations. The most heterogeneous cluster was the CS3, presenting features of both GC and mature B-cells (plasma cells and memory B-cells). Intriguingly, the cluster CS2 did not present any B-cell stage enrichment but instead a lymphatic endothelial cell (LEC) gene signature, suggesting wide transcriptomic or epigenetic changes (Fig. S26-S27). Finally, we aimed to confirmed the observed specific lymphoma's TME by using the CIBERSORTx derived immune deconvolution. The CS1 cluster was enriched in naive and memory B-cells but depleted in other immune populations; on the other hand, the other immune cold cluster (CS2) was only enriched in memory resting $CD4^+$ T-cells and monocytes. The CS3 cluster

was only enriched in inactivated macrophages M0. In line with the observed inflammatory phenotype observed using the Fges signatures, the CS4 cluster was enriched with $CD8^+$ T-cells, NK cells, activated macrophages M1, and memory activated $CD4^+$ T-cells (Fig. S28). A summary of these results is presented in Figure 3A for the FF cohort and in Fig. S29 for the FFPE cohort.

Immune deconvolution

VST transformed reads were used as input to determine immune cell types fractions using CIBERSORTx (B-mode batch correction; bootstrapping = 1000)[3].

TCR/BCR clonotypes

MiXCR (v3.0.3) was used to detect T-cell and B-cell clonotypes where concordant clonotypes were determined based on CDR3 amino acid sequence[77]. Gene usage, diversity estimation, and repertoire overlapping were estimated using immunarch (v0.6.6)[78]. The analyses were performed using the FF-RNA cohort (n=123).

Summary of results. To further characterize T-cells, BCR clonotypes, and the immunoglobulin (Ig) heavy-chain variable (V_H) and constant regions expression across clusters, we analyzed TCR/BCR-sequences in the complete set of FF-RNA (n=123). Concerning TCR-sequences, we found significantly higher T-cell clonotype diversity in the CS4 versus the others ($p < 0.05$), followed by CS3 and CS2, and then CS1 clusters (Fig. S30-S31). Regarding BCR-sequences analysis, we observed, even though not significantly different across clusters, the presence of BCR clones' diversity (Fig. S32) and $IgV_{H4\text{-}34}$ expression (Fig. S33) which further supports the presence of tumoral B-cells across clusters. However, there was a higher expression of IgM in the CS1/CS3 clusters compared to the CS4 which is characteristic of B-cells reentering the GC[79]. Additionally, the IgV_H regions more expressed in CS4 ($p < 0.05$) were the $V_{1\text{-}18}$, $V_{1\text{-}46,}$ and $V_{3\text{-}9}$ that have been associated with more differentiated B-cell stages; meanwhile, the naive-transition stage-related, $V_{3\text{-}43}$

and $V_{4\text{-}30\text{-}2}$ regions, were upregulated in CS3 (Fig. S33)[80].

Fusion transcripts expression

Fusion transcripts were then detected using arriba (v2.1.0) with previously generated STAR outputs[81]. Fusions were filtered according to the recommended "high fidelity" settings which correspond to having supporting split reads (split_reads1 + split_reads2 > max(1, discordant_mates/10)); plotting was done with circlize (v0.4.13)[82]. The analyses were performed using the FF-RNA cohort (n=123).

Summary of results. We proceeded with extending the fusion transcripts analysis across subtypes by analyzing the RNA-seq data from the whole FF cohort (n = 123). After filtering for "high fidelity" fusions, we identified some already described events involving *BCL6* (23.6%) and *ETV6* (14.6%) that happen mostly with the *IgH* super-enhancer (62.1% and 27.8%, respectively)[44,83]. Other frequent fusions involved *RNF213-ENDOV* (11.4%), *GRB2* (8.9%), and *KANSL1* (7.3%). Of note, we confirmed fusion enrichment of *CDKN2A/2B* and *CREBBP* in CS1 (p = 0.001 and 0.014, respectively); *OVOS* in CS3 (p = 0.004); and an additional of *CD274* (PD-L1) in CS2 (4/28 cases, p = 0.009, Fig. S34-S35).

Transcription factors regulon activity and master regulators

To determine TFs regulating the expression of subtype-specific genes, we applied DoRothEA to the gene expression matrix of each molecular PCNSL subtype (n_{CS1}=54; n_{CS2}=28; n_{CS3}=20; n_{CS4}=21) using the genes that are differentially expressed within each cluster by DESeq2 R software (BH adjusted P-value < 0.05 and $|Log_2FC|>1.5$)[27,84]. We selected the most reliable interactions from the curated collection of TF-targets provided by DoRothEA to compute TF activities, reported as TF regulon normalized enrichment scores (NESs), using the Wald statistic (retrieved from running DESeq2) on VIPER[85]. In VIPER, we set the eset.filter parameter to FALSE and consider five as the minimum number of targets allowed per regulon. We inferred the master regulators within each molecular subtype by

running RTN package (v2.16.0)[86] using the same gene expression matrices explained above and the regulators (TFs) obtained from DoRothEA. Regulator-target associations are identified using: I) mutual information (MI) which indicates whether or not a regulator is well informative of the status of a target gene and II) the direction of the association (positive or negative) evaluated by Spearman's correlation. Associations were filtered first by retaining only edges with a BH-adjusted p-value < 0.01 after permuting the MI matrix 1000 times; secondly by eliminating unstable interactions by 1000 times resamples' bootstrapping (consensus bootstrap > 95%); and finally, by removing indirect TF-target edges applying the Data Processing Inequality (DPI) filter with a 0 tolerance. The above computations do not consider associations' direction since it is later calculated. Master regulators (MRegs) analysis was performed using the tna.mra 'RTN' function which evaluates the overlap between each given regulon and the listed "hits" (Top 600 DE genes within each molecular subtype). MRegs were filtered to those having a BH adjusted p-value < 0.05[86].

Master regulator activity estimated by two-tailed GSEA

We performed a two-side GSEA analysis using the tni.gsea2 function in the RTN package with 1000 permutations and selecting the MRegs with a BH adjusted p-value < 0.05[86]. In general, for each MRegs, the approach divides the MRegs' targets into positive (A) and negative (B) that were previously defined using Spearman's correlation, then plots on top the DE (log_2-FC of all genes) observed when comparing the evaluated molecular subtype (CS1 to CS4) versus the other subtypes (this is called phenotype) in which genes are ranked from higher to lower log_2-FC values. The observed differential enrichment scores (dES) are the difference of the GSEA statistics in the ranked phenotype of A minus B where large positive dES indicates an induced regulon status while a large negative dES indicates the opposite case.

Summary of results. We found increased TF activity of *TEAD1* (in CS1 and CS2); *PRDM14* (CS1) which has roles as histone methyltransferase and leukemia initiator [87]; *IRF4,*

MYC, SPIB, and *PROX1* (CS2); *E2F1*, *SIX5*, and *IRF3* (CS3); and *STAT3*, *STAT1*, *NFKB1* and *TBX21* (CS4, Table S9, and Fig. S36-S39). Interestingly, given that the homeobox gene *PROX1* is a master regulator of LEC differentiation[88], together with the observed increased LEC gene signature and TF activity of *IRF4* and *MYC* in the CS2 cluster, support the idea of a phenotypic shift to more LEC-like in this PCNSL subtype. The next step was to integrate the association directionality for each target gene per TF by adding the expression fold changes and then to estimate the master regulator activity by two-tailed GSEA, as previously described[86]. CS2 exhibits induced regulon status for *MEIS1* whose targets include genes like *GNAI1* and *KCNJ2*; meanwhile, CS4 for *STAT1* which regulates important immune-related genes like *CXCL9, CD96, PD-LD2,* and *CD3G* (Table S10, and Fig. S40-S41).

Brain magnetic resonance imaging (MRI) analysis

MR images of PCNSL at time of first diagnosis and recurrence in sufficient quality were available for 90 of the PCNSL patients included in the validation cohort. Both T1-weighted images with contrast enhancement (CE) and fluid-attenuated inversion recovery (FLAIR)/T2-weighted axial images were reviewed for topographic tumor location. Image pre-processing encompassed N4 bias-field correction and linear co-registration using the open-source ANTs packages (https://stnava.github.io/ANTs/)[89]. The MRI were affine registered to the T1 sequence with ANTs and resampled to 1x1x3mm voxel size prior to segmentation. Tumor segmentation was performed semi-automatically for the contrast-enhancing portion using a region-growing segmentation algorithm implemented in ITK-SNAP v3.8 (by manually setting the parameters and initial seeds for two active contour algorithms)[90,91]. The tumor locations per cluster is given in Table S11 and represented in Figure 3B from the main article.

Methyl-seq

Sequencing

Methylation was performed by using the TruSeq Methyl Capture EPIC (Illumina)[92]. In total,

500 ng genomic DNA was used as input material, and the DNA was fragmented into around 150–200 bp by Covaris, followed by end-repair, 3' A-tailing, and adaptor ligation. Libraries were then pooled in groups of four in equal aliquot, on which two rounds of hybridization and capture using Illumina-optimized EPIC probe sets (covering >3.3 million targeted CpG sites), bisulfite conversion, and amplification were performed. Construction of DNA libraries and subsequent processing and DNA sequencing of paired-end reads (2 × 100nt reads) were performed according to the standard Illumina protocol using the Illumina Nova-seq 6000 sequencing system.

Alignment and quantification

Fastq data (n=68 from which: 27 were CS1; 17 CS2; 11 CS3; 9 CS4; 4 control normal blood samples) were aligned using Bismark (v0.23.1)[93], with the same reference genome (GRCh38) as for the WES data. The resulting aligned reads were used to generate methylation calls which excluded any duplicate calls from overlapping read ends of short inserts. Methylation percentages and read coverage for each CpG was calculated running the corresponding Bismark scripts[93]. Bismark coverage files were used as input for RnBeads (v2.12.2) analysis[94]. Sites that overlapped with SNPs, targeted sex chromosomes, had unreliable measurements (unknown chromosomes) and had exceptionally high/low coverage, were filtered. Imputation was performed by calculating the mean methylation level for each CpG site across all samples and replacing all missing values for this CpG site in individual samples with the mean across all samples. Imputation replaced a median of 16,540,067.5 missing values per sample by estimations. In total 21,934,824 out of 23,133,115 sites were retained. Differential methylation analysis (DMA) on gene promoters, CpG islands and regions was assessed by hierarchical linear models from the limma package, then p-values were adjusted by FDR. Gene promoters were defined as the 1,500 bases upstream and 500 bases downstream of the transcription start sites of corresponding genes and the differentially methylated promoters (DMP) were selected using the top 500 rank of RnBeads. Chromosome ends' beta-values were defined as the mean methylation

values within 4 Mb to the chromosome end considering chromosomes 1 to 22. Beta-values from all sites, CpG islands, promoters, genes and chromosome ends were compared across subtypes by using a two-sided Wilcoxon test (Fig. S42A-B). The top 1,000 most variable methylation sites were obtained using the median absolute deviation to observe the differences in a heatmap across subtypes and normal blood samples (Fig. S43).

Summary of results. The CS2 displayed higher hypermethylation globally, within promoters, and at chromosome ends (p=0.006, <0.001, and <0.001); however, the CS1 subtype presented higher methylation within CpG islands (p=0.001, Fig. S42A-B). Interestingly, an hypermethylator phenotype has been previously associated with a depleted TME in systemic DLBCL[70].

epiCMIT analysis

CpGs methylation β-values (per sample) were used as input to estimate de epiCMIT, epiCMIT-hyper, and epiCMIT-hypo scores according to the script "Estimate.epiCMIT.R" (https://github.com/Duran-FerrerM/Pan-B-cell-methylome), as previously described[95]. epiCMIT is the highest score between the hyper/hypo scores and reflects the relative accumulation of mitotic cell divisions of a particular sample, including the mitotic history associated with normal B-cell development as well as with malignant transformation and progression (including proliferation).

Summary of results. In line with the observed hypermethylation of CpG islands for CS1, this cluster exhibited the highest epiCMIT score ($p < 0.001$, Fig. S42C) which probably reflects its high proliferation rather than B-cell development. Moreover, since high epiCMIT scores were previously associated with worse OS in systemic DLBCL, we evaluated its clinical impact in PCNSL finding no association (log-rank p=0.42, Fig. S42D).

GO and LOLA enrichment analyses

GO enrichment analysis on DMP was performed using the GOstats R package as described in the RnBeads manual where enrichments were considered significant if having an FDR adjusted p-value <0.01[96]. Transcription factor binding sites (TFBs) enriched in the promoter regions across PCNSL molecular subtypes were identified by conducting enrichment analysis using the Bioconductor package LOLA (Locus Overlap enrichment Analysis, version 1.22.0)[97]. Specifically, all assessed promoters were used as background (universe) while the sets were divided for hypo or hypermethylated promoters per cluster type, all cell types in the LOLA Core database were included, and enrichments with a FDR-corrected p-value<0.01 were considered significant.

Summary of results. The complete list of enrichments is provided is Table S12-S13 and some are plotted in Fig. S44.

DNA methylation-transcriptome integrated analysis

For the integrated analysis of gene expression with DNA methylation, we either took the gene expression for all genes and all samples (for global analysis) or identified the DEGs (BH adjusted p-value < 0.05 and $|Log_2FC|>1.5$) per molecular subtype and then ran a Spearman's correlation analysis between the normalized expression count of each DEG with the methylation intensity (β-values) of its corresponding promoter or gene (covers methylation along the full gene). We then calculated the FDR using a permutation approach for each p-value.

Summary of results. Taking all PCNSL samples (n=64) and a total of 27,111 gens with available expression and methylation data, only 12.4% were correlated with their promoter methylation (FDR p.adjusted < 0.05 and Rho < 0). Interesting genes included *TERT, CD79A,* and *CD79B* (Rho=-0.50, -0.59, and -0.58 respectively; p.adjusted<0.001). Regarding the molecular subtypes, from the total 5,547 DE genes (95.4% down; 4.6% up) in CS1, 4,222 had available β-values from which only 0.4% were associated with their

promoter methylation (FDR p.adjusted < 0.05 and Rho < 0) and 0.2% with full gene methylation. For CS2 a total of 10,235 genes were DE (13.6% down; 86.4% up) from which 7,509 had β-values, furthermore, 0.2% were associated with the promoter and 0.1% with the full gene methylation. Regarding CS3, 194 genes were DE (71.6% down; 28.4% up) and 142 had β-values. For CS4, 401 genes were DE (32.9% down; 67.1% up) from which 339 had β-values. However, there was no correlation for any gene after p-value correction for the CS3/CS4 groups. Genes affected in CS2 include *CCL22*, *CD1C*, *CDCA7L*, and *CDK20.* Remarkably, *CCL22*, which is a chemoattractant for primary activated T-lymphocytes[98], is downregulated in the CS2 cluster and associated with gene hypermethylation. The complete list of spearman correlations is provided in Table S14.

Statistical analyses and figures

All statistical analyses were performed using the R statistical programming environment (version 4.0). Figures were generated using either base R or the ggplot2 library[99]. Differences in proportions and binary/categorical variables were calculated from two-sample Z-tests or Fisher's exact tests. Kruskal-Wallis test was used to test for a difference in distribution between three or more independent groups, and Mann Whitney U test was used for differences in distributions between two population groups unless otherwise noted. P-values were corrected for multiple comparisons using the Benjamini-Hochberg method when applicable. For heatmap representation (ComplexHeatmap package)[28], VST gene expression values were first quantile normalized and log2 transformed and then converted to Z-scores. Overall and progression-free survival analysis was assessed using log-rank Kaplan-Meier curves and univariate/multivariate Cox proportional hazards regression modeling. For each omic-level data (Fusions, CNV, TME, BCR/TCR clonotypes, RNA-seq, and mutations) the Least Absolute Shrinkage and Selection Operator (LASSO)-Cox method was implemented to select variables associated with the OS. We used an L1 penalized (alpha = 1) model where the performance score and age were kept unpenalized and then evaluated each prognostic model by Harrell's concordance index (C-index) in both the

discovery (n=85) and validation FFPE cohort (n = 93; only TME and RNA-seq data)[100].

Data and Code Availability

Genomic data will be deposited at the European Genome-phenome Archive (EGA; http://www.ebi.ac.uk/ega/), which is hosted by the EBI.

The RBraLymP algorithm is publicly accessible at https://github.com/iS4i4S/PCNSL-RBraLymP. All other materials are available upon request from the authors.

References

1. Houillier C, Soussain C, Ghesquières H, et al. Management and outcome of primary CNS lymphoma in the modern era: An LOC network study. Neurology 2020;94(10):e1027–39.
2. Wilms T, Khan G, Coates PJ, et al. No evidence for the presence of Epstein-Barr virus in squamous cell carcinoma of the mobile tongue. PLOS ONE 2017;12(9):e0184201.
3. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 2019;37(7):773–82.
4. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12(5):453–7.
5. Lopez G, Egolf LE, Giorgi FM, Diskin SJ, Margolin AA. *svpluscnv*: analysis and visualization of complex structural variation data. Bioinformatics 2021;37(13):1912–4.
6. Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31(13):3812–4.
7. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 2012;40(W1):W452–7.
8. Sunyaev S. Prediction of deleterious human alleles. Hum Mol Genet 2001;10(6):591–7.
9. Ramensky V. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002;30(17):3894–900.
10. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics 2018;19(1):71–86.
11. Meng C, Helm D, Frejno M, Kuster B. moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. J Proteome Res 2016;15(3):755–65.
12. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat Commun 2018;9(1):4453.
13. Chalise P, Fridley BL. Integrative clustering of multi-level ‘omic data based on non-negative matrix factorization algorithm. PLOS ONE 2017;12(5):e0176278.
14. Monti S. [No title found]. Mach Learn 2003;52(1/2):91–118.
15. Hoadley KA, Yau C, Wolf DM, et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell 2014;158(4):929–44.
16. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics 2019;35(18):3348–56.
17. Nguyen H, Shrestha S, Draghici S, Nguyen T. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics 2019;35(16):2843–6.
18. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11(3):333–7.
19. Wu D, Wang D, Zhang MQ, Gu J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genomics 2015;16(1):1022.
20. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
21. Lu X, Meng J, Zhou Y, Jiang L, Yan F. *MOVICS*: an R package for multi-omics integration and visualization in cancer subtyping. Bioinformatics 2021;36(22–23):5539–41.
22. Xiong B, Yang Y, Fineis FR, Wang J-P. DegNorm: normalization of generalized transcript degradation improves accuracy in RNA-seq analysis. Genome Biol 2019;20(1):75.
23. Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High

Throughput Sequence Data [Internet]. 2010 [cited 2020 Mar 12]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
24. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21.
25. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 2011;17(1):10.
26. Wang L, Nie J, Sicotte H, et al. Measure transcript integrity using RNA-seq data. BMC Bioinformatics 2016;17(1):58.
27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
28. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016;
29. Hoshida Y. Nearest Template Prediction: A Single-Sample-Based Flexible Class Prediction with Confidence Assessment. PLoS ONE 2010;5(11):e15543.
30. Broad Institute. Picard Tools [Internet]. 2018. Available from: http://broadinstitute.github.io/picard/
31. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio [Internet] 2013 [cited 2022 Feb 18];Available from: http://arxiv.org/abs/1303.3997
32. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297–303.
33. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma 2013;43:11.10.1-11.10.33.
34. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43(5):491–8.
35. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods 2018;15(8):591–4.
36. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22(3):568–76.
37. Wilm A, Aw PPK, Bertrand D, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res 2012;40(22):11189–201.
38. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 2012;28(3):311–7.
39. Fang LT, Afshar PT, Chhibber A, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biol 2015;16(1):197.
40. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17(1):122.
41. Jonsson P. facetsSuite [Internet]. Available from: https://github.com/mskcc/facets-suite
42. Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res 2018;28(11):1747–56.
43. Martincorena I, Raine KM, Gerstung M, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 2017;171(5):1029-1041.e21.
44. Chapuy B, Roemer MGM, Stewart C, et al. Targetable genetic features of primary testicular and primary central nervous system lymphomas. Blood 2016;127(7):869–81.
45. Schmitz R, Wright GW, Huang DW, et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. N Engl J Med 2018;378(15):1396–407.
46. Chapuy B, Stewart C, Dunford AJ, et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. Nat Med 2018;24(5):679–90.
47. Verdin IH, Akdemir KC, Ramazzotti D, et al. Pan-cancer landscape of AID-related

mutations, composite mutations and its potential role in the ICI response [Internet]. Bioinformatics; 2021 [cited 2022 Feb 16]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.06.26.447715
48. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature 2013;500(7463):415–21.
49. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. Nature 2020;578(7793):94–101.
50. Shinde J, Bayard Q, Imbeaud S, et al. Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. Bioinformatics [Internet] 2018 [cited 2020 Mar 12];Available from: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty388/4996591
51. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci 2004;101(12):4164–9.
52. Wang S, Li H, Song M, et al. Copy number signature analyses in prostate cancer reveal distinct etiologies and clinical outcomes [Internet]. Genetic and Genomic Medicine; 2020 [cited 2020 Nov 23]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.04.27.20082404
53. Yin X, Bi R, Ma P, et al. Multiregion whole-genome sequencing depicts intratumour heterogeneity and punctuated evolution in ovarian clear cell carcinoma. J Med Genet 2020;57(9):605–9.
54. PCAWG Evolution & Heterogeneity Working Group, PCAWG Consortium, Gerstung M, et al. The evolutionary history of 2,658 cancers. Nature 2020;578(7793):122–8.
55. 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–e131.
56. Hundal J, Carreno BM, Petti AA, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med 2016;8(1):11.
57. Griffithlab. VAtools [Internet]. 2018. Available from: https://vatools.org
58. Fleri W, Salimi N, Vita R, Peters B, Sette A. Immune Epitope Database and Analysis Resource [Internet]. In: Encyclopedia of Immunobiology. Elsevier; 2016 [cited 2020 Mar 19]. p. 220–4.Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780123742797060045
59. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst 2018;7(1):129-132.e4.
60. Shao XM, Bhattacharya R, Huang J, et al. High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. Cancer Immunol Res 2020;8(3):396–408.
61. Schmidt J, Smith AR, Magnin M, et al. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. Cell Rep Med 2021;2(2):100194.
62. Talevich E, Shain AH. CNVkit-RNA: Copy number inference from RNA-Sequencing data [Internet]. Bioinformatics; 2018 [cited 2020 May 19]. Available from: http://biorxiv.org/lookup/doi/10.1101/408534
63. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. 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.
64. de Reyniès A, Jaurand M-C, Renier A, et al. Molecular Classification of Malignant Pleural Mesothelioma: Identification of a Poor Prognosis Subgroup Linked to the Epithelial-to-Mesenchymal Transition. Clin Cancer Res 2014;20(5):1323–34.
65. 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 2003;100(17):9991–6.
66. Reddy A, Zhang J, Davis NS, et al. Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. Cell 2017;171(2):481-494.e15.

67. Orenbuch R, Filip I, Comito D, Shaman J, Pe'er I, Rabadan R. arcasHLA: high-resolution HLA typing from RNAseq. Bioinformatics 2020;36(1):33–40.
68. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS J Integr Biol 2012;16(5):284–7.
69. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 2013;14(1):7.
70. Kotlov N, Bagaev A, Revuelta MV, et al. Clinical and Biological Subtypes of B-cell Lymphoma Revealed by Microenvironmental Signatures. Cancer Discov 2021;11(6):1468–89.
71. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. Cell 2015;160(1–2):48–61.
72. Yusufova N, Kloetgen A, Teater M, et al. Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. Nature 2021;589(7841):299–305.
73. Dersh D, Phelan JD, Gumina ME, et al. Genome-wide Screens Identify Lineage- and Tumor-Specific Genes Modulating MHC-I- and MHC-II-Restricted Immunosurveillance of Human Lymphomas. Immunity 2021;54(1):116-131.e10.
74. Wright GW, Huang DW, Phelan JD, et al. A Probabilistic Classification Tool for Genetic Subtypes of Diffuse Large B Cell Lymphoma with Therapeutic Implications. Cancer Cell 2020;37(4):551-568.e14.
75. Milpied P, Cervera-Marzal I, Mollichella M-L, et al. Human germinal center transcriptional programs are de-synchronized in B cell lymphoma. Nat Immunol 2018;19(9):1013–24.
76. Schubert M, Klinger B, Klünemann M, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun 2018;9(1):20.
77. Bolotin DA, Poslavsky S, Davydov AN, et al. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol 2017;35(10):908–11.
78. ImmunoMind Team. immunarch: An R Package for Painless Analysis of Large-Scale Immune Repertoire Data [Internet]. 2019. Available from: https://doi.org/10.5281/zenodo.3367200
79. Dogan I, Bertocci B, Vilmont V, et al. Multiple layers of B cell memory with different effector functions. Nat Immunol 2009;10(12):1292–9.
80. Wu Y-C, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood 2010;116(7):1070–8.
81. Uhrig S, Ellermann J, Walther T, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res 2021;31(3):448–60.
82. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics 2014;30(19):2811–2.
83. Bruno A, Labreche K, Daniau M, et al. Identification of novel recurrent ETV6-IgH fusions in primary central nervous system lymphoma. Neuro-Oncol 2018;20(8):1092–100.
84. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res 2019;29(8):1363–75.
85. Alvarez MJ, Shen Y, Giorgi FM, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet 2016;48(8):838–47.
86. Castro MAA, de Santiago I, Campbell TM, et al. Regulators of genetic risk of breast cancer identified by integrative network analysis. Nat Genet 2016;48(1):12–21.
87. Dettman EJ, Simko SJ, Ayanga B, et al. Prdm14 initiates lymphoblastic leukemia after expanding a population of cells resembling common lymphoid progenitors. Oncogene 2011;30(25):2859–73.
88. Jalkanen S, Salmi M. Lymphatic endothelial cells of the lymph node. Nat Rev Immunol 2020;20(9):566–78.
89. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation

of ANTs similarity metric performance in brain image registration. NeuroImage 2011;54(3):2033–44.
90. Caselles V, Kimmel R, Sapiro G. [No title found]. Int J Comput Vis 1997;22(1):61–79.
91. Song Chun Zhu, Yuille A. Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE Trans Pattern Anal Mach Intell 1996;18(9):884–900.
92. Masser DR, Stanford DR, Freeman WM. Targeted DNA Methylation Analysis by Next-generation Sequencing. J Vis Exp 2015;(96):52488.
93. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 2011;27(11):1571–2.
94. Müller F, Scherer M, Assenov Y, et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. Genome Biol 2019;20(1):55.
95. Duran-Ferrer M, Clot G, Nadeu F, et al. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. Nat Cancer 2020;1(11):1066–81.
96. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics 2007;23(2):257–8.
97. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics 2016;32(4):587–9.
98. Wiedemann GM, Knott MML, Vetter VK, et al. Cancer cell-derived IL-1α induces CCL22 and the recruitment of regulatory T cells. OncoImmunology 2016;5(9):e1175794.
99. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org
100. Goeman JJ. L_1 Penalized Estimation in the Cox Proportional Hazards Model. Biom J 2009;NA-NA.

Supplementary Tables

Table of contents

Table	Description
S1*	PCNSL FF and FFPE cohorts clinical data and immunohistochemical panel
S2	Summarization of clinical features within clusters
S3*	Top 100 subtype specific biomarkers
S4*	Mutations and Indels summary
S5*	Driver genes
S6*	AID targets and affected samples
S7*	CNA and RNA correlations differential expression analysis
S8*	Distinct genetic signatures within PCNSL subtypes
S9*	DoRothEA NES activities
S10*	Two-tailed GSEA master regulator activity
S11*	FFPE MRI locations
S12*	Promoter's methylation GO enrichment analysis
S13*	Promoter's methylation LOLA analysis
S14*	DNA methylation-transcriptome spearman correlations

*See online links to Excel spreadsheets

Supplemental Tables Legends

(*See online links to Excel spreadsheets for other than Table S2)

Table S1. Clinical data, including median overall survival, median progression free survival, and number of patients who received intensive chemotherapy with autologous stem cell rescue in the FF and FFPE cohorts (S1A). Immunohistochemical panel of antibodies used for PCNSL diagnosis that includes the company, clone and dilution (S1B).

Table S2. Summary of clinical features such as OS, PFS, IC-ASCR, sex, age, immunogenic neoantigens, TMB, fraction of c-AID mutations, Memorial Sloan Kettering Cancer Center (MSKCC) classification, Karnofsky index (IK), and purity across the detected PCNSL significant clusters (CS1 to CS4).

Table S3. List of top 100 unique upregulated biomarkers (no overlapping across subtypes) used for extrapolating molecular classification using one level omic-data (RNA-seq) in either FF or FFPE samples. Biomarkers were identified for each molecular subtype (Deseq2 package v1.32.0) with a threshold of adjusted p-value < 0.05.

Table S4. Summary per sample in the FF cohort (n=115) of mutations and indels obtained from WES.

Table S5. Results from running ddNdScv to obtain driver genes in PCNSL samples (n=115) were ultra-hypermutator samples and extremely mutated genes per sample were filtered to avoid losing sensitivity. Genes were considered as drivers (n=466) if having q-values < 0.001 (Benjamini-Hochberg's multiple testing correction of p-values; column named as "qglobal_cv").

Table S6. AID off-target mutated genes and frequency and number of samples affected in

the FF cohort (n=115).

Table S7. Differential expression analysis (Deseq2 package v1.32.0) resulting from comparing samples having a specific focal deletion or amplification versus those samples not having the event (Related to Fig. S21).

Table S8. Number of samples in each cluster (CS1 to CS4) presenting an event (mutation, fusion, amplification or deletion) and their corresponding p-value using a Fisher one-sided test within-cluster vs outside-cluster. P values were corrected by FDR ("q.value" column). Related to Figure 2C from the main article.

Table S9. Normalized enrichment scores (NESs) for the TF activities computed by DoRothEA for each molecular cluster.

Table S10. Differential enrichment scores to measure the master regulator activity estimated by two-tailed GSEA.

Table S11. Number of tumors per cluster associated with different MRI location (e.g., splenium, insula, etc.) or characteristic (important tumor size, meningeal infiltration, etc.). P-values were calculated by a one-sided Fisher test.

Table S12. Go enrichment results from using the differentially methylated promoters in clusters CS1 (S12A), CS2 (S12B), CS3 (S12C), and CS4 (S12D).

Table S13. LOLA enrichment results from using the differentially methylated promoters in clusters CS1 (S13A), CS2 (S13B), CS3 (S13C), and CS4 (S13D).

Table S14. Spearman's Rho, p-value and FDR adjusted p-value resulting from DNA

methylation (promoters or gene body) and transcriptome expression correlations using all samples (n=64; S14A) or samples corresponding to the different clusters (S14B-14I).

Table S2. Summary of clinical features within clusters

	Level	CS1	CS2	CS3	CS4	p-value	Test
n		39	19	18	9		
censor (%)	0	22 (56.4)	6 (31.6)	3 (16.7)	8 (88.9)	<0.001	exact
	1	17 (43.6)	13 (68.4)	15 (83.3)	1 (11.1)		
OS (days) (median [IQR])		799.10 [407.17, 1948.95]	549.00 [253.15, 1630.23]	419.38 [187.57, 510.11]	2037.40 [603.90, 2049.60]	<0.001	log-rank
PFS (days) (median [IQR])		366.00 [183.00, 610.00]	146.40 [103.70, 750.30]	181.48 [100.65, 472.75]	1281.00 [506.30, 2037.4]	0.045	log-rank
PFS status (censor %)	0	11 (31.4)	3 (20.0)	4 (22.2)	6 (66.7)	0.98	exact
	1	24 (68.6)	12 (80.0)	14 (77.8)	3 (33.3)		
	NA	4	4	0	0		
IC-ASCR (%)	0	31 (79.5)	14 (87.5)	15 (88.2)	9 (100)	0.323	exact
	1	8 (20.5)	2 (12.5)	2 (11.8)	0 (0)		
	NA	0	3	1	0		
Sex (%)	F	27 (69.2)	10 (52.6)	5 (27.8)	5 (55.6)	0.031	exact
	M	12 (30.8)	9 (47.4)	13 (72.2)	4 (44.4)		
Immunogenic Neoags (median [IQR])		9.00 [4.75, 20.50]	10.00 [3.50, 12.25]	6.00 [3.75, 11.75]	14.00 [6.00, 24.00]	0.439	nonnorm
TMB per Mb (median [IQR])		3.86 [2.92, 5.85]	1.72 [0.19, 3.97]	3.86 [2.99, 7.41]	4.30 [3.77, 4.84]	0.012	nonnorm
KPS (%)	0	2 (5.3)	1 (5.3)	1 (5.6)	0 (0.0)	0.768	exact
	10	1 (2.6)	0 (0.0)	0 (0.0)	0 (0.0)		
	20	1 (2.6)	0 (0.0)	0 (0.0)	0 (0.0)		
	30	2 (5.3)	0 (0.0)	2 (11.1)	0 (0.0)		
	40	6 (15.8)	3 (15.8)	1 (5.6)	0 (0.0)		
	50	3 (7.9)	3 (15.8)	4 (22.2)	0 (0.0)		
	60	5 (13.2)	4 (21.1)	2 (11.1)	1 (11.1)		
	70	10 (26.3)	2 (10.5)	3 (16.7)	3 (33.3)		
	80	5 (13.2)	3 (15.8)	5 (27.8)	4 (44.4)		
	90	3 (7.9)	3 (15.8)	0 (0.0)	1 (11.1)		
Fraction c-AID mutations		0.08 [0.06, 0.10]	0.06 [0.04, 0.09]	0.09 [0.07, 0.10]	0.08 [0.06, 0.09]	0.249	nonnorm

(median [IQR])							
CNV sigGroup (%)	CNV_Sig1	20 (51.3)	5 (26.3)	4 (22.2)	2 (22.2)	0.002	exact
	CNV_Sig2	1 (2.6)	8 (42.1)	2 (11.1)	2 (22.2)		
	CNV_Sig3	5 (12.8)	0 (0.0)	1 (5.6)	2 (22.2)		
	CNV_Sig4	13 (33.3)	6 (31.6)	11 (61.1)	3 (33.3)		
Age		67.21 ± 12.70	64.74 ± 15.66	67.83 ± 12.39	70.67 ± 6.75	0.523	
MSKCC (%)	Class1	4 (10.8)	2 (10.5)	4 (23.5)	0 (0.0)	0.144	exact
	Class2	17 (45.9)	11 (57.9)	5 (29.4)	8 (88.9)		
	Class3	16 (43.2)	6 (31.6)	8 (47.1)	1 (11.1)		
Purity (median [IQR])		0.78 [0.66, 0.88]	0.48 [0.29, 0.70]	0.70 [0.63, 0.78]	0.60 [0.39, 0.77]	0.010	nonnorm
KPS binary (%)	<70	20 (52.6)	11 (57.9)	10 (55.6)	2 (22.2)	0.326	exact
	>=70	18 (47.4)	8 (42.1)	8 (44.4)	7 (77.8)		
Age binary (%)	<median	21 (53.8)	11 (57.9)	7 (38.9)	2 (22.2)	0.250	exact
	>=median	18 (46.2)	8 (42.1)	11 (61.1)	7 (77.8)		

Note: OS, overall survival. PFS, progression-free survival. MSKCC, Memorial Sloan Kettering Cancer Center. KPS, Karnofsky Performance Status. CNV sigGroup, copy number variation signature group. Purity was obtained from WES estimations using copy-number inferences. Exact, Fisher-exact test. Nonnorm, Wilcoxon-test.

Supplementary Figures

Table of contents

S25	PROGENY signatures comparison across PCNSL subtypes
S26	B-cell differentiation signatures comparison across PCNSL subtypes
S27	B-cell TFs signatures comparison across PCNSL subtypes
S28	CIBERSORTx derived immune cells comparison across PCNSL subtypes
S29	Phenotypic distinctions in the FFPE validation cohort
S30	B-cell and T-cell clonotype's diversity
S31	Chao diversity for TCR clones
S32	Chao diversity for BCR clones
S33	Immunoglobulin heavy variable segments comparison across the PCNSL subtypes
S34	Fusion transcripts expression across the PCNSL molecular subtypes
S35	CDKN2A/2B fusion transcript schematic illustration
S36	CS1 DoRothEA TFs NES activities
S37	CS2 DoRothEA TFs NES activities
S38	CS3 DoRothEA TFs NES activities
S39	CS4 DoRothEA TFs NES activities
S40	Master regulators targets in the CS groups
S41	Two-way GSEA activity for *MEIS1* and *STAT1*
S42	Mean methylation and epiCMIT differences across CS groups
S43	Methylation heatmap across CS groups and controls
S44	LOLA enrichment results across PCNSL molecular subtypes

Supplemental Figures Legends

Figure S1. Prediction of optimal cluster number of multi-omics clusters by cluster prediction index and Gap-statistics in the multi-omic FF cohort (n=85).

Figure S2. Consensus heatmap based on the 10 integrative clustering algorithms to refine the clusters (CS1 to CS4) where each of the 10 algorithms uses cluster of clusters analysis to integrate six levels of omic information (y-axis) in the order: i) mRNA expression (2,087 variables), ii) mutations (31 variables), iii) CNA (40 variables), iv) fusion transcripts 43 variables), v) TCR/BCR clonotypes (19 variables), and vi) immune cell fractions (22 variables).

Figure S3. Quantification of sample similarity using silhouette score based on the consensus ensembles result. Each line in the y-axis represents a sample.

Figure S4. Panel A shows a forest plot of a Cox model of the cluster's impact in OS after adjustment by IK (<70 or ≥ 70) and age (<median or ≥ median). Panel B shows Kaplan-Meier estimates of PFS among patients belonging to each cluster that resulted from the Consensus cluster of clusters analysis. Panel C shows a forest plot of a Cox model of the cluster's impact in PFS after adjustment by IK (<70 or ≥ 70) and age (<median or ≥ median). All the results were obtained using the multi-omic FF cohort (n=85).

Figure S5. Bar plot of fraction genome altered among clusters. Error bars indicate 95% binomial confidence intervals. Asterisks denote significant difference by Kruskal-Wallis rank sum test for multiple subtypes.

Figure S6. Panel A shows the alignment efficiency comparison between 3' FFPE RNA-seq

and FF RNA-seq paired data (coming from the same samples). Asterisks denote significant differences by Wilcoxon test. Panel B shows the RNA-seq reads coverage over the gene body in 3' FFPE RNA-seq and FF RNA-seq paired data where numbers in parenthesis indicates the transcript integrety number (0-100 where lower values indicates more degradation). Panel C shows the different mismatch profiles (number of mismatches in y-axis) across the read position 5' to 3' (x-axis) in the 3' FFPE RNA-seq and FF RNA-seq paired data.

Figure S7. Spearman correlations of VST gene read counts between FF (x-axis) and FFPE (y-axis) pairs with either the uncorrected FFPE data (Panel A) or the degradation corrected FFPE data (Panel B). "R" indicates Spearman Rho values and "p" the associated p-values.

Figure S8. Panel A shows the cell-of-origin (COO, ABC or GCB subtypes) assignment of each tumor using either the uncorrected FFPE data (left) or corrected FFPE data (right). Panel B shows a correlation heatmap between FF-FFPE pairs. Panel C shows PCA plots for FF-FFPE pairs.

Figure S9. Consistency heatmap between samples (n=85) assigned by using consensus multi-omic clustering (CMOIC, y-axis) or transcriptomic-based PAM method (x-axis). Numbers in the diagonal indicate the samples correctly assigned by PAM which was 100% for CS1 (39/39), 84% for CS2 (16/19), 83% for CS3 (15/18), and 100% for CS4 (9/9). Accuracy of assignment evaluated by Cohen's kappa coefficient.

Figure S10. Panel A shows Kaplan-Meier estimates of OS among patients belonging to each cluster that resulted from assigning patients using only transcriptomic data by the PAM method. Panel B shows a forest plot of a Cox model of the cluster's impact in OS after adjustment by IK (<70 or ≥ 70) and age (<median or ≥ median). These results were obtained using the FF-RNA cohort (n=123).

Figure S11. Heatmap of NTP method in the FFPE cohort (n=93) using the top 100 subtype-specific upregulated biomarkers to predict the CS groups.

Figure S12. Panel A shows a forest plot of a Cox model of the cluster's impact in OS after adjustment by IK (<70 or ≥ 70) and age (<median or ≥ median). Panel B shows Kaplan-Meier estimates of PFS among patients belonging to each cluster that resulted from the Consensus cluster of clusters analysis. Panel C shows a forest plot of a Cox model of the cluster's impact in PFS after adjustment by IK (<70 or ≥ 70) and age (<median or ≥ median). All the results were obtained using the FFPE cohort (n=93).

Figure S13. Frequency of mutations according to different classifications (e.g. missense, nonsense, etc), types (TNP, SNP, etc) or class (T>G, T>A, etc). Frequency of top mutated genes (bottom right). Results from WES of the PCNSL FF cohort (n=115).

Figure S14. Oncoplot of the top 27 driver genes in PCNSL (n=115) where each column represents a sample and each row a gene. Top bar plot shows the TMB per sample filled according to mutation type (missense, nonsense, splice site, frameshift, multihit, or other). Right bar plot shows the number of affected samples filled according to mutation type.

Figure S15. Boxplot comparison of the fraction of c-AID mutations versus APOBEC mutations detected in the FF cohort. Dash lines indicate the median of each variable. P-value computed by Wilcoxon test.

Figure S16. Rainfall plots of the AID mutations' distribution across chromosomes on PCNSL samples as a function of logarithmic (10 scale) genomic distance. Black points represent C to G mutations and red points C to T mutations. Middle barplot shows the sum of mutations (density) across chromosomes. Top barplot shows examples of c-AID off-targets followed by

the frequency of affected samples where genes are marked as bold if they are driver genes. Genes were marked as “Kataegic targets” if they had consecutive mutations with an average mutational distance ≤1 Kb within the same patient.

Figure S17. Boxplot comparison of the proportion of mutations attributed to different mutational processes when occurring at clonal-early times versus clonal-late times. P-values calculated by Wilcoxon matched-pairs signed rank test for paired data (per sample).

Figure S18. Boxplot comparison of the proportion of mutations attributed to different mutational processes when occurring at clonal times versus subclonal times. P-values calculated by Wilcoxon matched-pairs signed rank test for paired data (per sample).

Figure S19. GISTIC 2.0 results of significantly recurring (events with a q-value of less than 0.1) focal amplifications (red) and deletions (blue) where the chromosomes are plotted in the x-axis and the GISTIC-scores (G-score) are plotted in the y-axis. Genes affected for each focal event are annotated followed by the percentage of altered samples (n=108). Genes are underlined if they are driver genes or in bold if they are transcriptionally affected by the focal event (Related to Fig. S21 and Table S7). FC, fold-change.

Figure S20. GISTIC mirror plots showing the significantly recurring CNAs in the PCNSL (n =108, right side) and DLBCL (n=296, left side) cohorts. Global arm deletions (Panel A), focal arm deletions (Panel B), global arm amplifications (Panel C), and focal arm amplifications (Panel D). Y-axis corresponds to chromosome number and x-axis to significant association as -log10 q-value.

Figure S21. Volcano plot showing the gene expression up-regulation among samples (FF cohort with available RNA-seq data, n=80) with copy gain and gene expression down-

regulation among samples with copy loss for specific focal CNAs. Genes with FDR<0.1 and absolute fold change >1.0 were considered significantly affected. Complete list of affected genes is given in Table S7.

Figure S22. Boxplot comparison of the mutational processes' (SBS, DBS, and ID) signature activity across the PCNSL molecular subtypes (n=85). P-values calculated by Wilcoxon test.

Figure S23. Boxplot comparison of the Fges score of different gene signatures across the PCNSL molecular subtypes (n = 85). P-values calculated by Wilcoxon test. Fges score calculated by GSVA analysis. LEC, lymphatic endothelial cell; VEC, vascular endothelial cell; CAF, cancer-associated fibroblasts; FRC, fibroblastic reticular cells; ECM, extracellular matrix; FDC, follicular dendritic cells; TFH, follicular helper T-cells, TIL, tumor infiltrating lymphocytes.

Figure S24. Boxplot comparison of the GSVA score of different oncogenic signatures across the PCNSL molecular subtypes (n=85). P-values calculated by Wilcoxon test.

Figure S25. Boxplot comparison of the PROGENy's score of different oncogenic signatures across the PCNSL molecular subtypes (n=85). P-values calculated by Wilcoxon test.

Figure S26. Boxplot comparison of the GSVA score of different B-cell differentiation signatures across the PCNSL molecular subtypes (n=85). P-values calculated by Wilcoxon test.

Figure S27. Boxplot comparison of the GSVA score of different B-cell TFs signatures across the PCNSL molecular subtypes (n = 85). P-values calculated by Wilcoxon test.

Figure S28. Boxplot comparison of the CIBERSORTx derived immune cells' proportions across the PCNSL molecular subtypes (n=85). P-values calculated by Wilcoxon test.

Figure S29. Heatmap with either gene signature activity (measured by GSVA) or immune cell proportions (CiberSortx deconvoluted) across molecular subtypes in the FFPE cohort (n=93). LEC, lymphatic endothelial cell; VEC, vascular endothelial cell; CAF, cancer-associated fibroblasts; FRC, fibroblastic reticular cells; ECM, extracellular matrix; FDC, follicular dendritic cells; TFH, follicular helper T-cells, TIL, tumor infiltrating lymphocytes.

Figure S30. Boxplot comparison of the number of unique B-cell (top) or T-cell (bottom) clonotypes across the PCNSL molecular subtypes (FF-RNA cohort, n=123). P-values calculated by Wilcoxon test.

Figure S31. Boxplot comparison of the Chao diversity for TCR clones across the PCNSL molecular subtypes (FF-RNA cohort, n=123). P-values calculated by Wilcoxon test.

Figure S32. Boxplot comparison of the Chao diversity for BCR clones across the PCNSL molecular subtypes (FF-RNA cohort, n=123). P-values calculated by Wilcoxon test.

Figure S33. Boxplot comparison of the gene usage fraction of different immunoglobulin heavy variable segments across the PCNSL molecular subtypes (FF-RNA cohort, n=123). P-values calculated by Wilcoxon test.

Figure S34. Circos plot showing the fusion transcripts expression across the PCNSL molecular subtypes (FF-RNA cohort, n=123). Fusions found to be enriched (according to Figure 2C from the main article) or present in at least 5 different samples were annotated

along with their frequency within each CS group. Fusions present in more than 8 samples are marked in bold.

Figure S35. Examples of *CDKN2A* and *CDKN2B* fusions detected.

Figure S36. TF regulon normalized enrichment scores (NESs) in the TFs found in the CS1 group by DoRothEA analysis.

Figure S37. TF regulon normalized enrichment scores (NESs) in the TFs found in the CS2 group by DoRothEA analysis.

Figure S38. TF regulon normalized enrichment scores (NESs) in the TFs found in the CS3 group by DoRothEA analysis.

Figure S39. TF regulon normalized enrichment scores (NESs) in the TFs found in the CS4 group by DoRothEA analysis.

Figure S40. Detected master regulators and their targets for each CS group.

Figure S41. Master regulator activity for MEIS1 (CS2 cluster, Panel A) or STAT1 (CS4 cluster, Panel B) estimated by two-tailed GSEA (left) and the regulated targets (right). The top plot indicates the DE (log2-FC of all genes) observed comparing the evaluated molecular subtype (CS2 or CS4) versus the other subtypes (called phenotype) in which genes are ranked from higher to lower log2-FC values. The bar beneath the phenotype shows red marks for activated and blue marks for repressed members of the MRegs. The GSEA plots show the running enrichment score for positive (red line) and negative (blue line) targets where the differential enrichment score (dES) indicates an induced or repressed MRegs status.

Figure S42. Boxplot comparison of either CS2 versus other clusters (Panel A) or across molecular subtypes (Panel B) using the mean methylation levels (mean beta-values) globally, on CpG islands, on promoters, on gene body ("genes") and at chromosome ends (4 Mb). Panel C shows a boxplot comparison across molecular subtypes of the epiCMIT, epiCMIT-hyper, and epiCMIT-hypo scores. Panel D shows Kaplan-Meier estimates of OS among patients belonging to each molecular PCNSL subtype using the median epiCMIT score to stratify. P-values calculated by Wilcoxon test. Results from methylation data of the PCNSL FF cohort (n=64).

Figure S43. Heatmap of the 1,000 most variable methylated sites across PCNSL molecular subtypes and blood control samples. Each row represents a site and each column a patient sample. The level of DNA methylation (beta-value) is represented with the color scale as depicted. Hierarchical clustering (Euclidean distance followed by complete-linkage agglomeration algorithm) was used to group rows and samples within subtypes. Results from methylation data of the PCNSL FF cohort (n=64 PCNSL and 4 controls).

Figure S44. LOLA analysis results of the targets of the enriched transcription factor binding sites retrieved using the differentially methylated promoters across PNCSL molecular subtypes. Y-axis denotes the log(p-value) and x-axis the target. Top annotation indicates the database from which the targets were found.

Figure S1

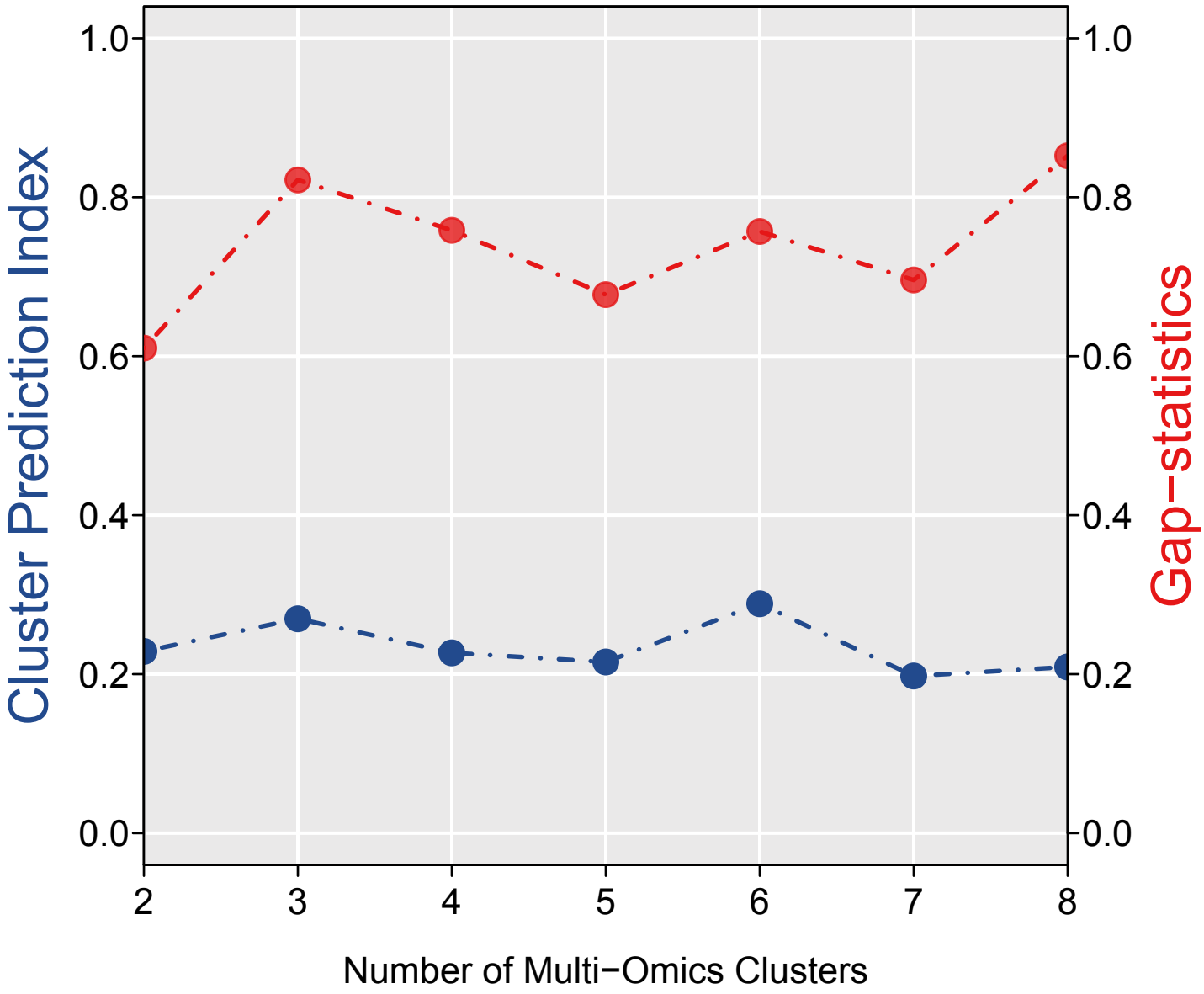

Cluster Prediction Index
Gap−statistics
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
7
8
Number of Multi−Omics Clusters

Figure S2

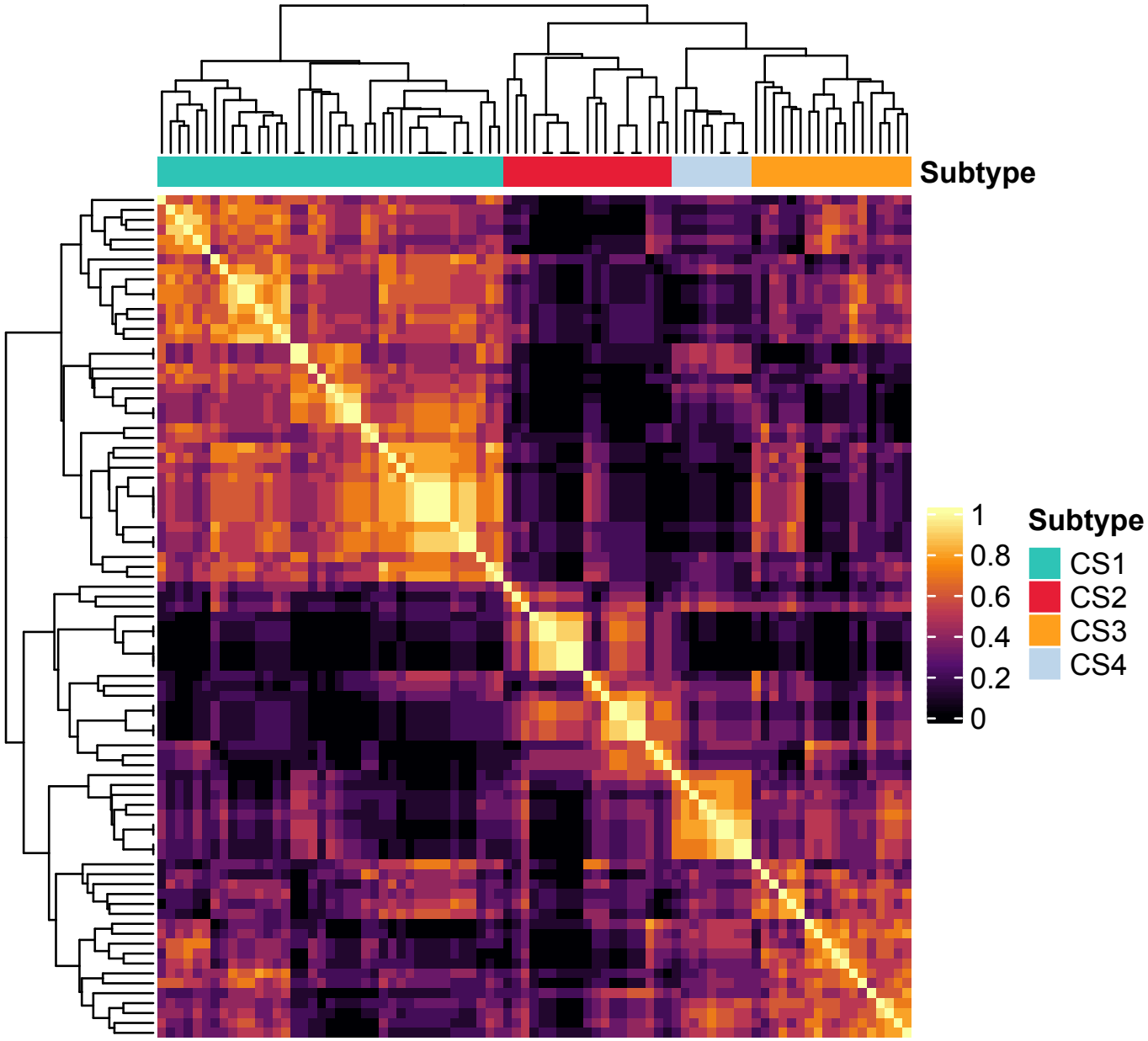


Figure S3

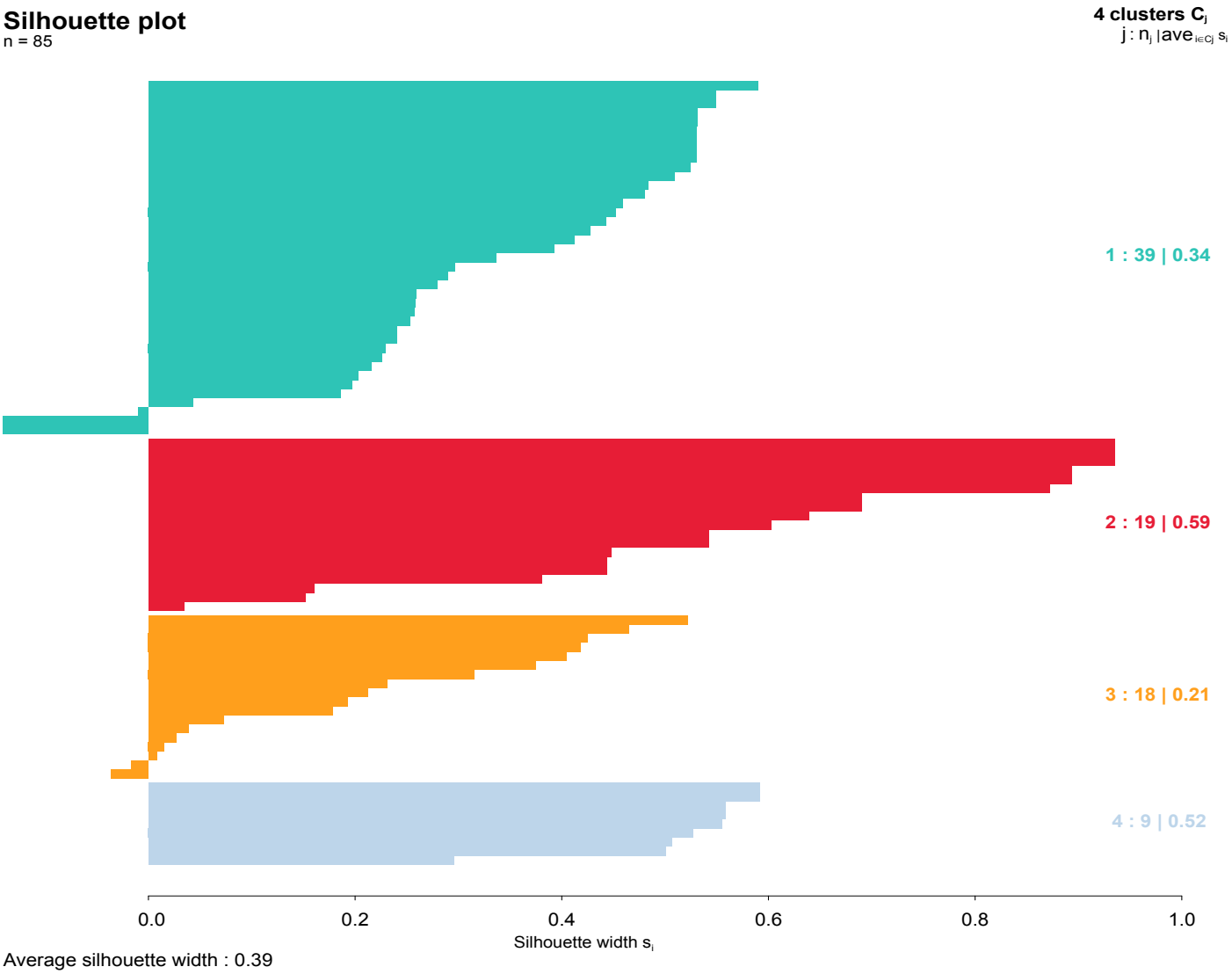

Silhouette plot
n = 85
4 clusters C_j
$j : n_j \mid ave_{i \in Cj}\, s_i$
1 : 39 | 0.34
2 : 19 | 0.59
3 : 18 | 0.21
4 : 9 | 0.52
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.39

Figure S4

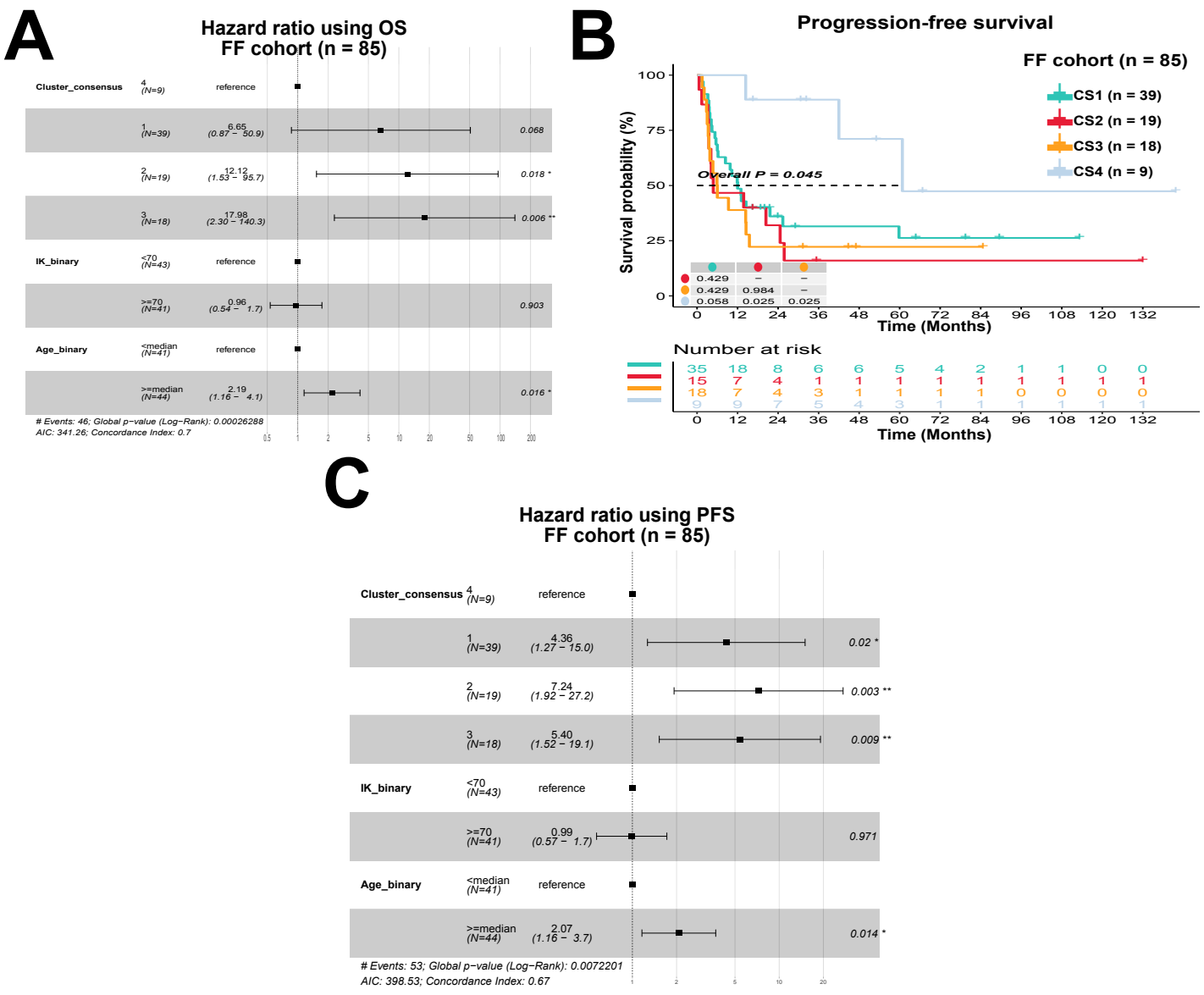

A
Hazard ratio using OS
FF cohort (n = 85)
Cluster_consensus
4 (N=9) reference
1 (N=39) 6.65 (0.87 – 50.9) 0.068
2 (N=19) 12.12 (1.53 – 95.7) 0.018 *
3 (N=18) 17.98 (2.30 – 140.3) 0.006 **
IK_binary
<70 (N=43) reference
>=70 (N=41) 0.96 (0.54 – 1.7) 0.903
Age_binary
<median (N=41) reference
>=median (N=44) 2.19 (1.16 – 4.1) 0.016 *
Events: 46; Global p−value (Log−Rank): 0.00026288
AIC: 341.26; Concordance Index: 0.7
B
Progression-free survival
FF cohort (n = 85)
CS1 (n = 39)
CS2 (n = 19)
CS3 (n = 18)
CS4 (n = 9)
Overall P = 0.045
Survival probability (%)
Time (Months)
Number at risk
C
Hazard ratio using PFS
FF cohort (n = 85)
Cluster_consensus
4 (N=9) reference
1 (N=39) 4.36 (1.27 – 15.0) 0.02 *
2 (N=19) 7.24 (1.92 – 27.2) 0.003 **
3 (N=18) 5.40 (1.52 – 19.1) 0.009 **
IK_binary
<70 (N=43) reference
>=70 (N=41) 0.99 (0.57 – 1.7) 0.971
Age_binary
<median (N=41) reference
>=median (N=44) 2.07 (1.16 – 3.7) 0.014 *
Events: 53; Global p−value (Log−Rank): 0.0072201
AIC: 398.53; Concordance Index: 0.67

Figure S5

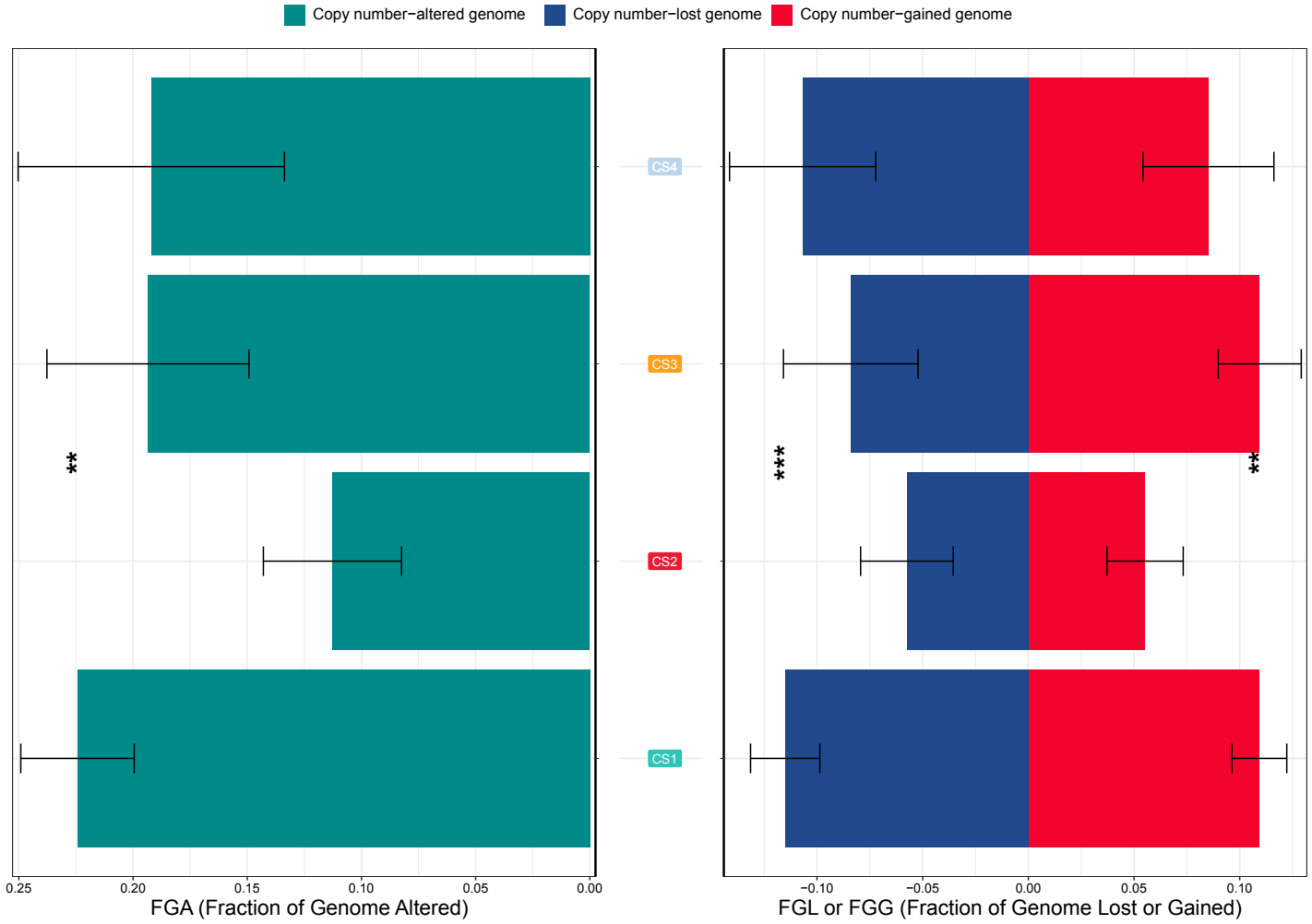

Copy number−altered genome
Copy number−lost genome
Copy number−gained genome
CS4
CS3
CS2
CS1
**
***
**
0.25
0.20
0.15
0.10
0.05
0.00
FGA (Fraction of Genome Altered)
−0.10
−0.05
0.00
0.05
0.10
FGL or FGG (Fraction of Genome Lost or Gained)

Figure S6

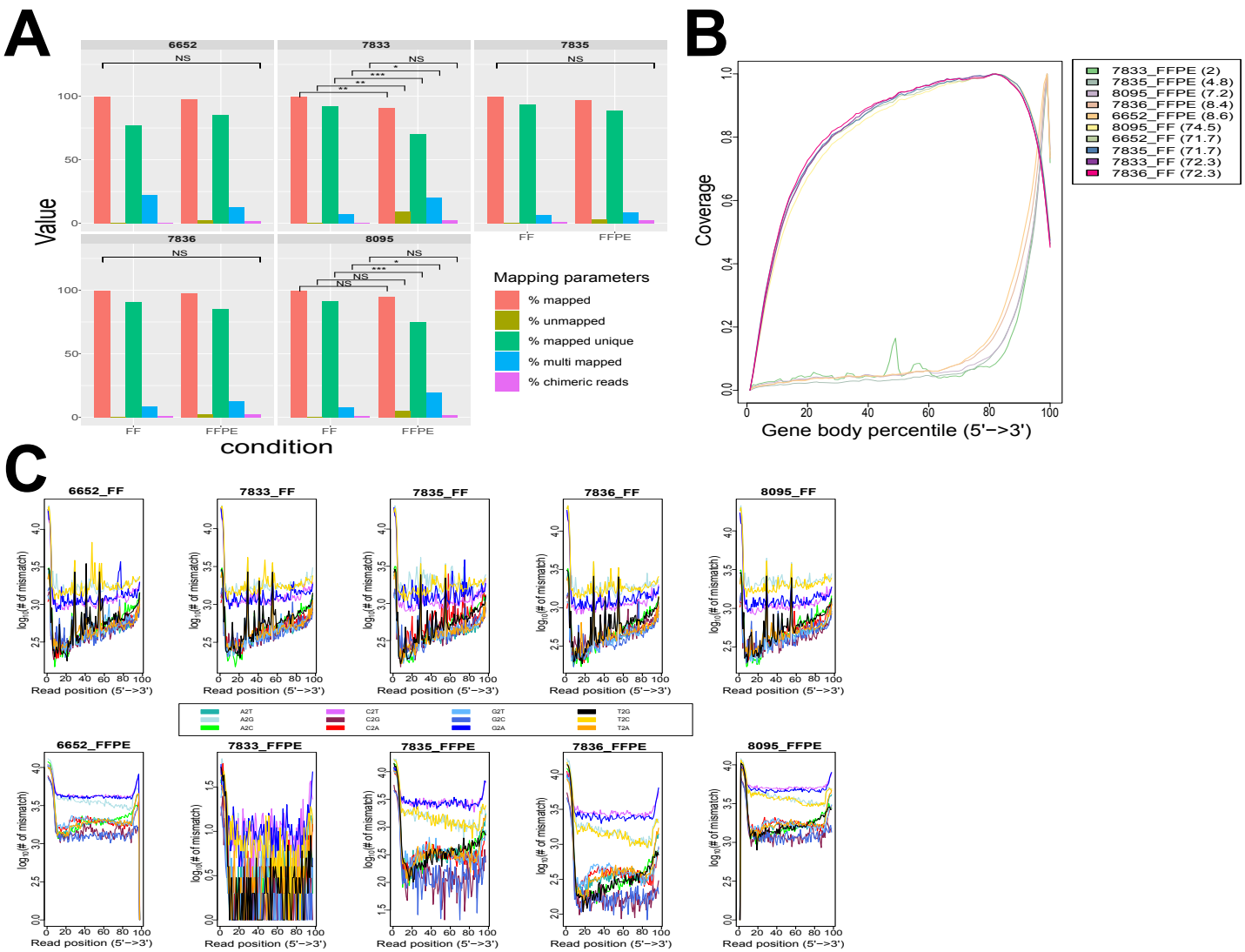

A
6652
7833
7835
7836
8095
NS
Value
condition
FF
FFPE
Mapping parameters
% mapped
% unmapped
% mapped unique
% multi mapped
% chimeric reads
B
Coverage
Gene body percentile (5'->3')
7833_FFPE (2)
7835_FFPE (4.8)
8095_FFPE (7.2)
7836_FFPE (8.4)
6652_FFPE (8.6)
8095_FF (74.5)
6652_FF (71.7)
7835_FF (71.7)
7833_FF (72.3)
7836_FF (72.3)
C
6652_FF
7833_FF
7835_FF
7836_FF
8095_FF
6652_FFPE
7833_FFPE
7835_FFPE
7836_FFPE
8095_FFPE
Read position (5'->3')
A2T
A2G
A2C
C2T
C2G
C2A
G2T
G2C
G2A
T2G
T2C
T2A

Figure S7

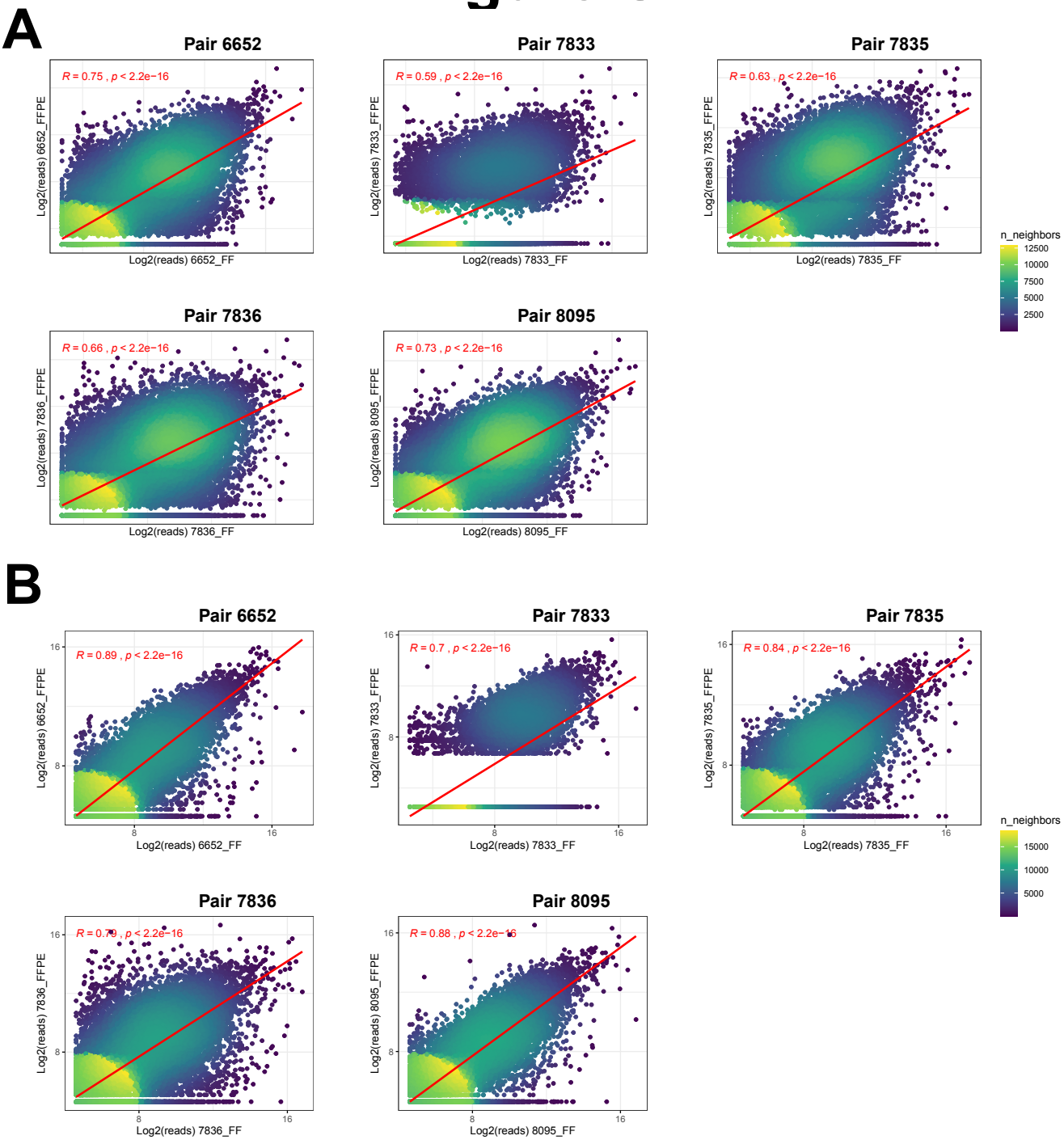

A
Pair 6652
R = 0.75 , p < 2.2e−16
Log2(reads) 6652_FFPE
Log2(reads) 6652_FF
Pair 7833
R = 0.59 , p < 2.2e−16
Log2(reads) 7833_FFPE
Log2(reads) 7833_FF
Pair 7835
R = 0.63 , p < 2.2e−16
Log2(reads) 7835_FFPE
Log2(reads) 7835_FF
n_neighbors
12500
10000
7500
5000
2500
Pair 7836
R = 0.66 , p < 2.2e−16
Log2(reads) 7836_FFPE
Log2(reads) 7836_FF
Pair 8095
R = 0.73 , p < 2.2e−16
Log2(reads) 8095_FFPE
Log2(reads) 8095_FF
B
Pair 6652
R = 0.89 , p < 2.2e−16
Log2(reads) 6652_FFPE
Log2(reads) 6652_FF
Pair 7833
R = 0.7 , p < 2.2e−16
Log2(reads) 7833_FFPE
Log2(reads) 7833_FF
Pair 7835
R = 0.84 , p < 2.2e−16
Log2(reads) 7835_FFPE
Log2(reads) 7835_FF
n_neighbors
15000
10000
5000
Pair 7836
R = 0.79 , p < 2.2e−16
Log2(reads) 7836_FFPE
Log2(reads) 7836_FF
Pair 8095
R = 0.88 , p < 2.2e−16
Log2(reads) 8095_FFPE
Log2(reads) 8095_FF

Figure S8

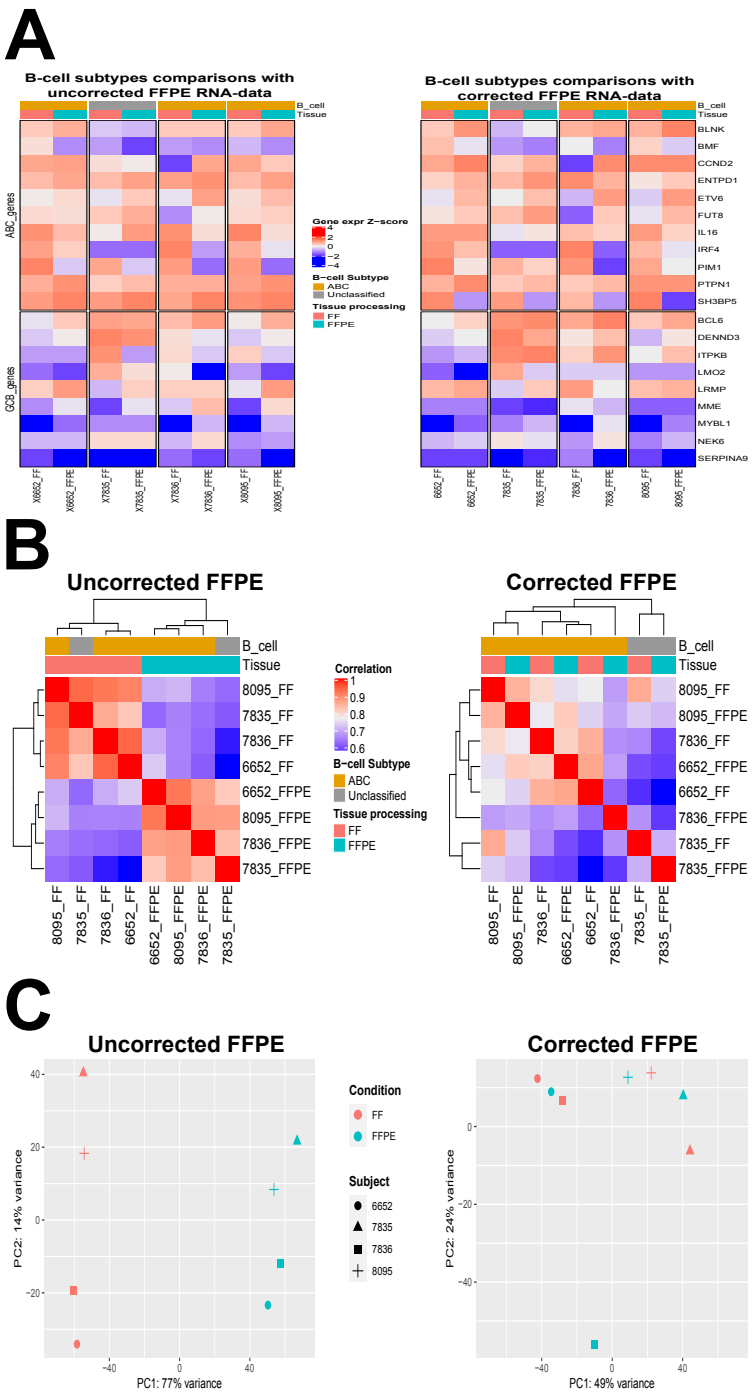

A
B-cell subtypes comparisons with uncorrected FFPE RNA-data
B-cell subtypes comparisons with corrected FFPE RNA-data
ABC_genes
GCB_genes
Gene expr Z-score
B-cell Subtype
ABC
Unclassified
Tissue processing
FF
FFPE
BLNK
BMF
CCND2
ENTPD1
ETV6
FUT8
IL16
IRF4
PIM1
PTPN1
SH3BP5
BCL6
DENND3
ITPKB
LMO2
LRMP
MME
MYBL1
NEK6
SERPINA9
B
Uncorrected FFPE
Corrected FFPE
Correlation
8095_FF
7835_FF
7836_FF
6652_FF
6652_FFPE
8095_FFPE
7836_FFPE
7835_FFPE
C
Uncorrected FFPE
Corrected FFPE
Condition
FF
FFPE
Subject
6652
7835
7836
8095
PC1: 77% variance
PC2: 14% variance
PC1: 49% variance
PC2: 24% variance

Figure S9

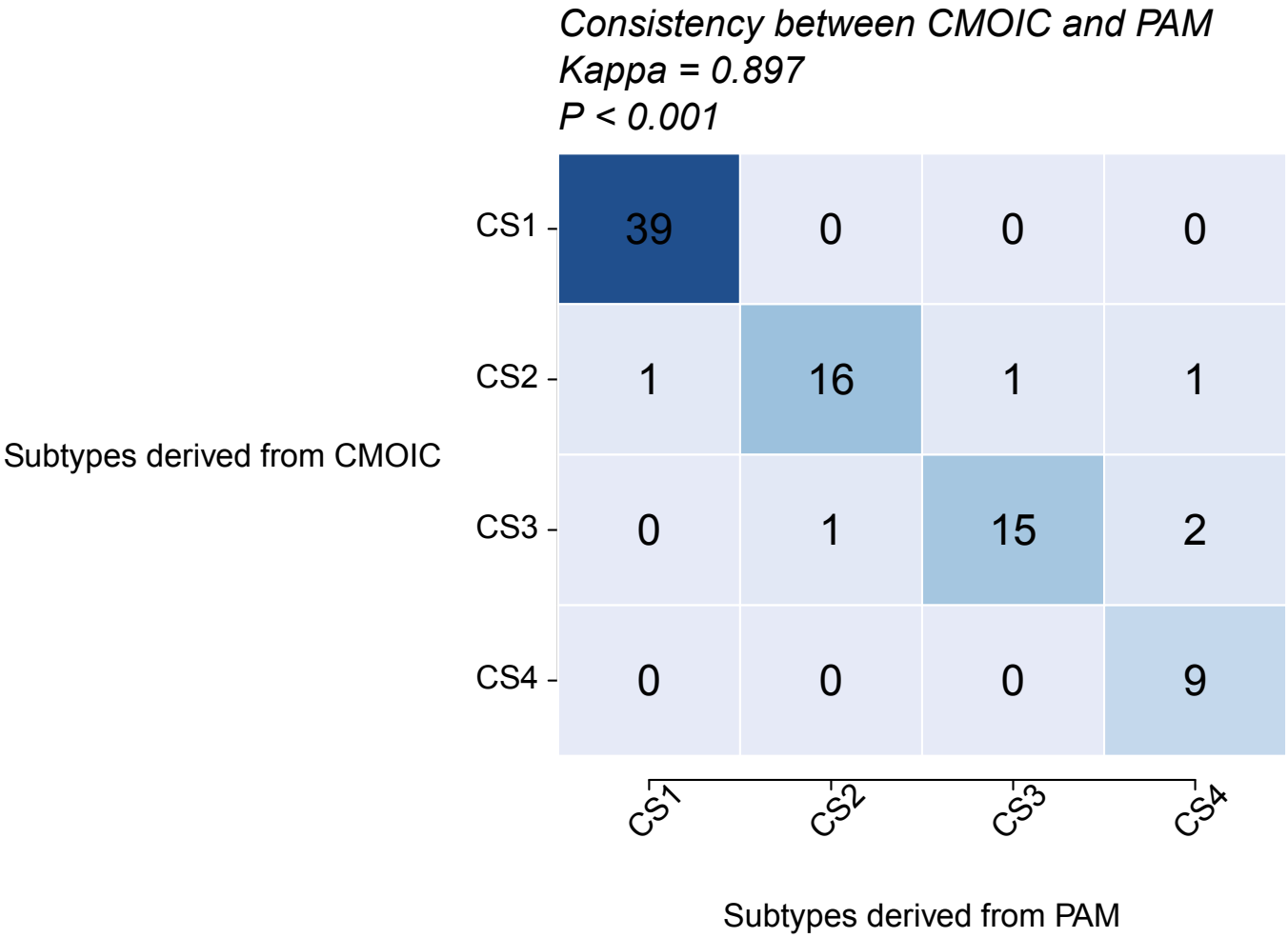

Consistency between CMOIC and PAM
Kappa = 0.897
P < 0.001
CS1
CS2
CS3
CS4
39
0
0
0
1
16
1
1
0
1
15
2
0
0
0
9
Subtypes derived from CMOIC
CS1
CS2
CS3
CS4
Subtypes derived from PAM

Figure S10

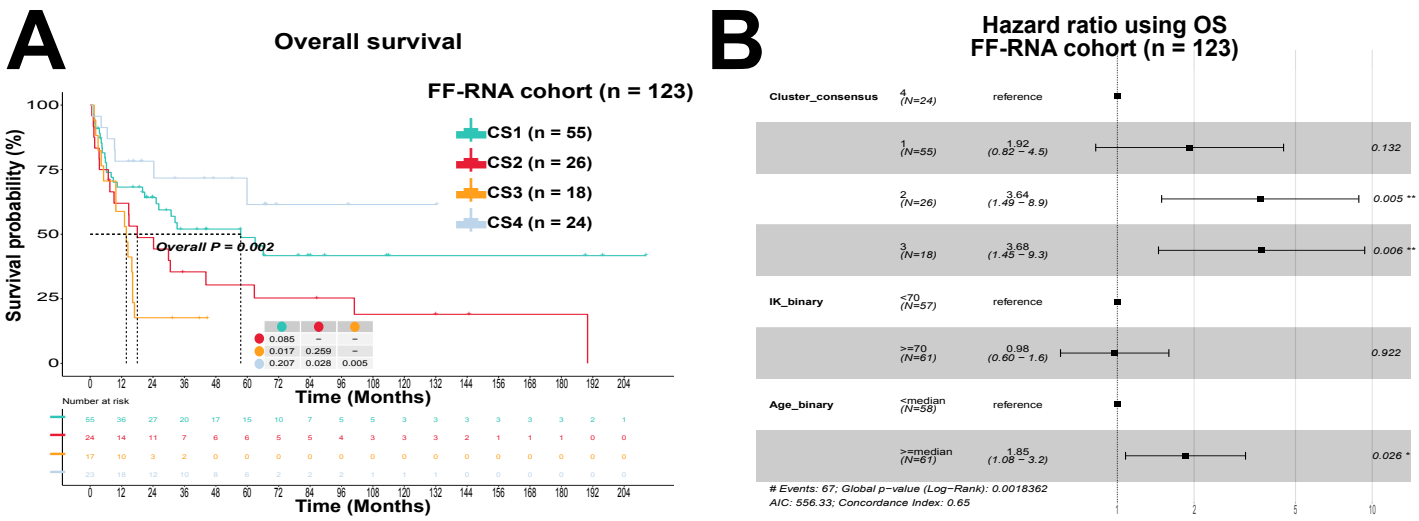

A
Overall survival
FF-RNA cohort (n = 123)
CS1 (n = 55)
CS2 (n = 26)
CS3 (n = 18)
CS4 (n = 24)
Survival probability (%)
Overall P = 0.002
Time (Months)
Number at risk
B
Hazard ratio using OS
FF-RNA cohort (n = 123)
Cluster_consensus
reference
IK_binary
Age_binary
Events: 67; Global p−value (Log−Rank): 0.0018362
AIC: 556.33; Concordance Index: 0.65

Figure S11

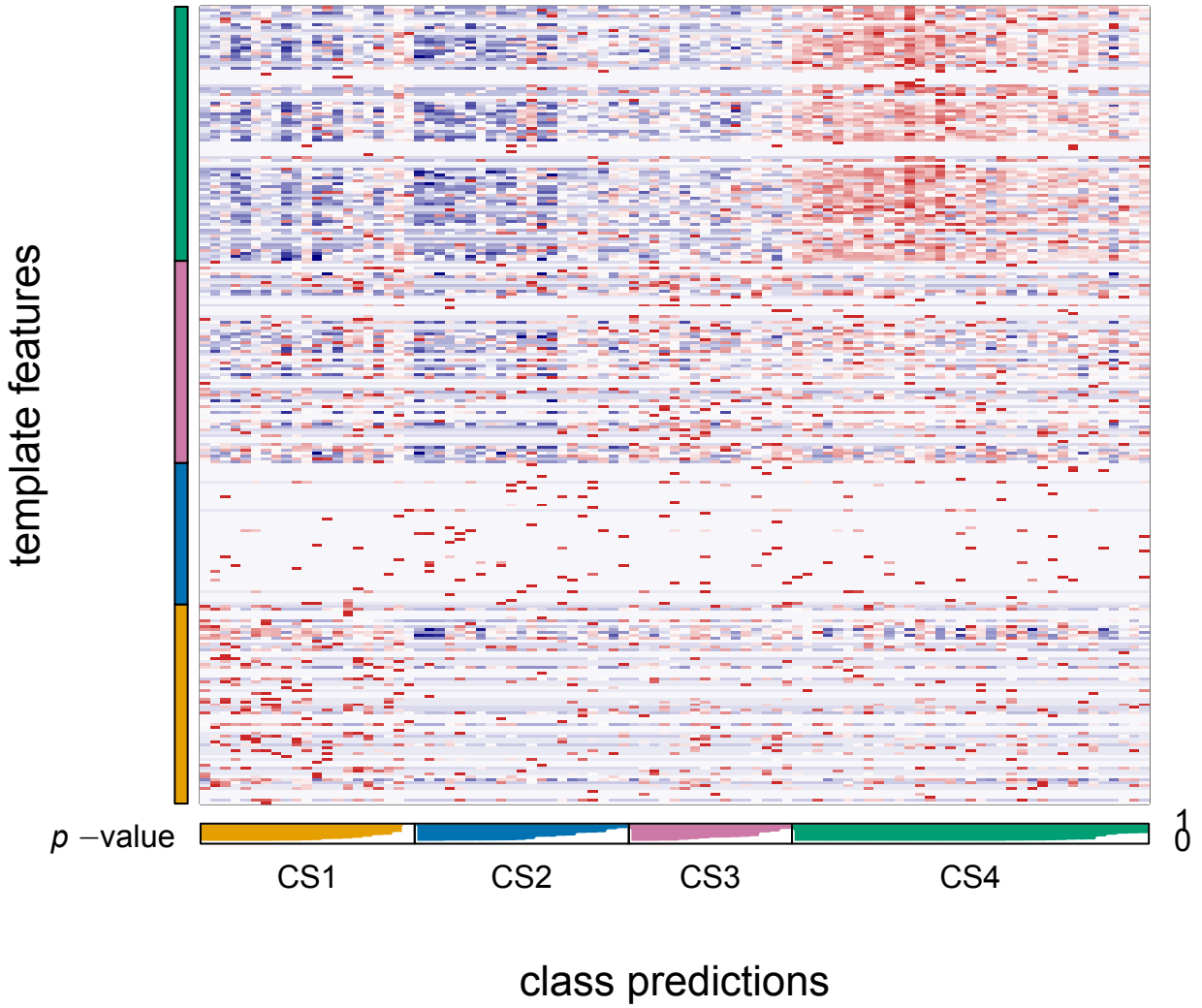

template features
p −value
1
0
CS1
CS2
CS3
CS4
class predictions

Figure S12

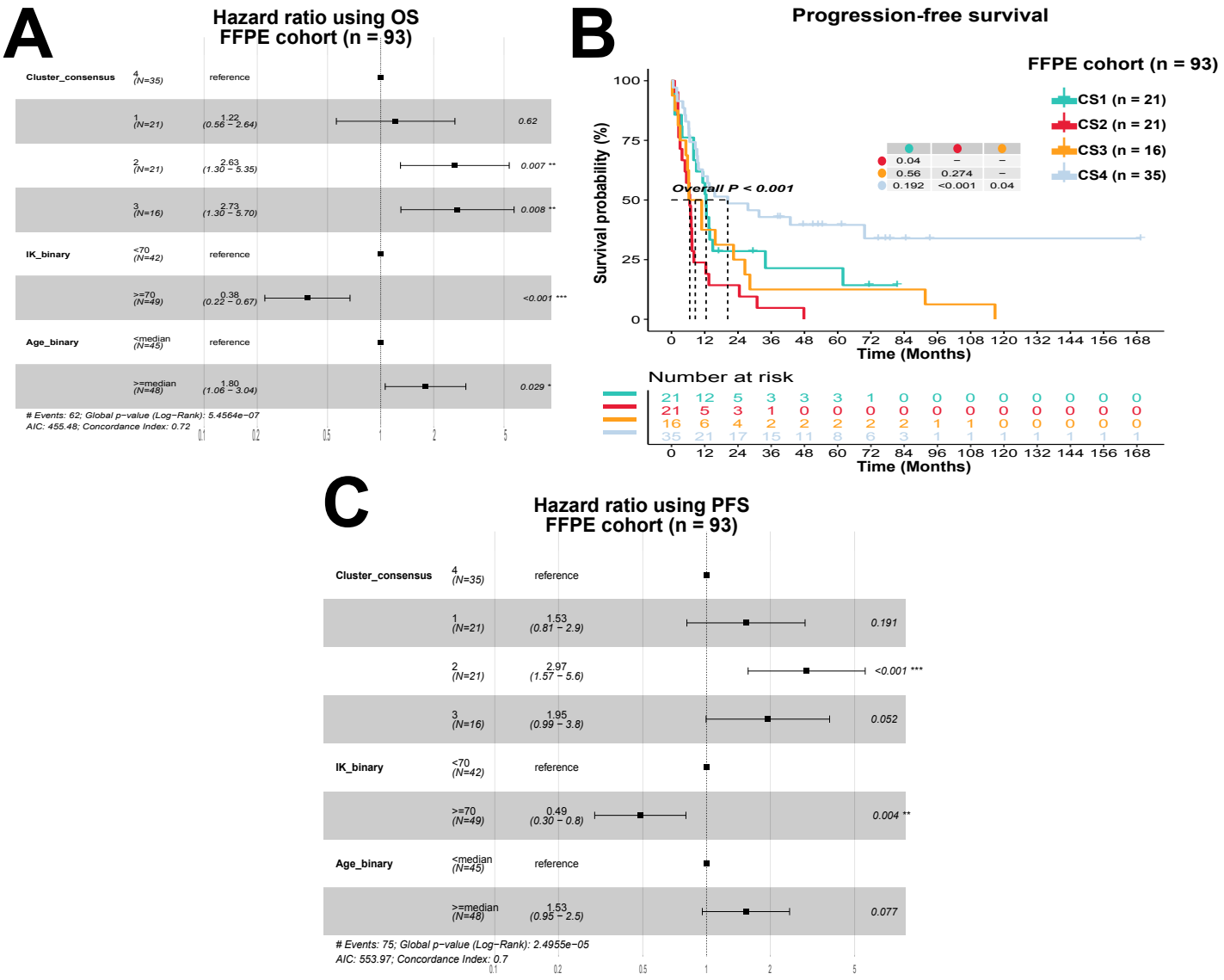

A
Hazard ratio using OS
FFPE cohort (n = 93)
Cluster_consensus
4 (N=35) reference
1 (N=21) 1.22 (0.56 – 2.64) 0.62
2 (N=21) 2.63 (1.30 – 5.35) 0.007 **
3 (N=16) 2.73 (1.30 – 5.70) 0.008 **
IK_binary
<70 (N=42) reference
>=70 (N=49) 0.38 (0.22 – 0.67) <0.001 ***
Age_binary
<median (N=45) reference
>=median (N=48) 1.80 (1.06 – 3.04) 0.029 *
Events: 62; Global p−value (Log−Rank): 5.4564e−07
AIC: 455.48; Concordance Index: 0.72
0.1 0.2 0.5 1 2 5
B
Progression-free survival
FFPE cohort (n = 93)
CS1 (n = 21)
CS2 (n = 21)
CS3 (n = 16)
CS4 (n = 35)
Survival probability (%)
Overall P < 0.001
0.04 – –
0.56 0.274 –
0.192 <0.001 0.04
Time (Months)
Number at risk
21 12 5 3 3 3 1 0 0 0 0 0 0 0 0
21 5 3 1 0 0 0 0 0 0 0 0 0 0 0
16 6 4 2 2 2 2 2 1 1 0 0 0 0 0
35 21 17 15 11 8 6 3 1 1 1 1 1 1 1
Time (Months)
C
Hazard ratio using PFS
FFPE cohort (n = 93)
Cluster_consensus
4 (N=35) reference
1 (N=21) 1.53 (0.81 – 2.9) 0.191
2 (N=21) 2.97 (1.57 – 5.6) <0.001 ***
3 (N=16) 1.95 (0.99 – 3.8) 0.052
IK_binary
<70 (N=42) reference
>=70 (N=49) 0.49 (0.30 – 0.8) 0.004 **
Age_binary
<median (N=45) reference
>=median (N=48) 1.53 (0.95 – 2.5) 0.077
Events: 75; Global p−value (Log−Rank): 2.4955e−05
AIC: 553.97; Concordance Index: 0.7
0.1 0.2 0.5 1 2 5

Figure S13

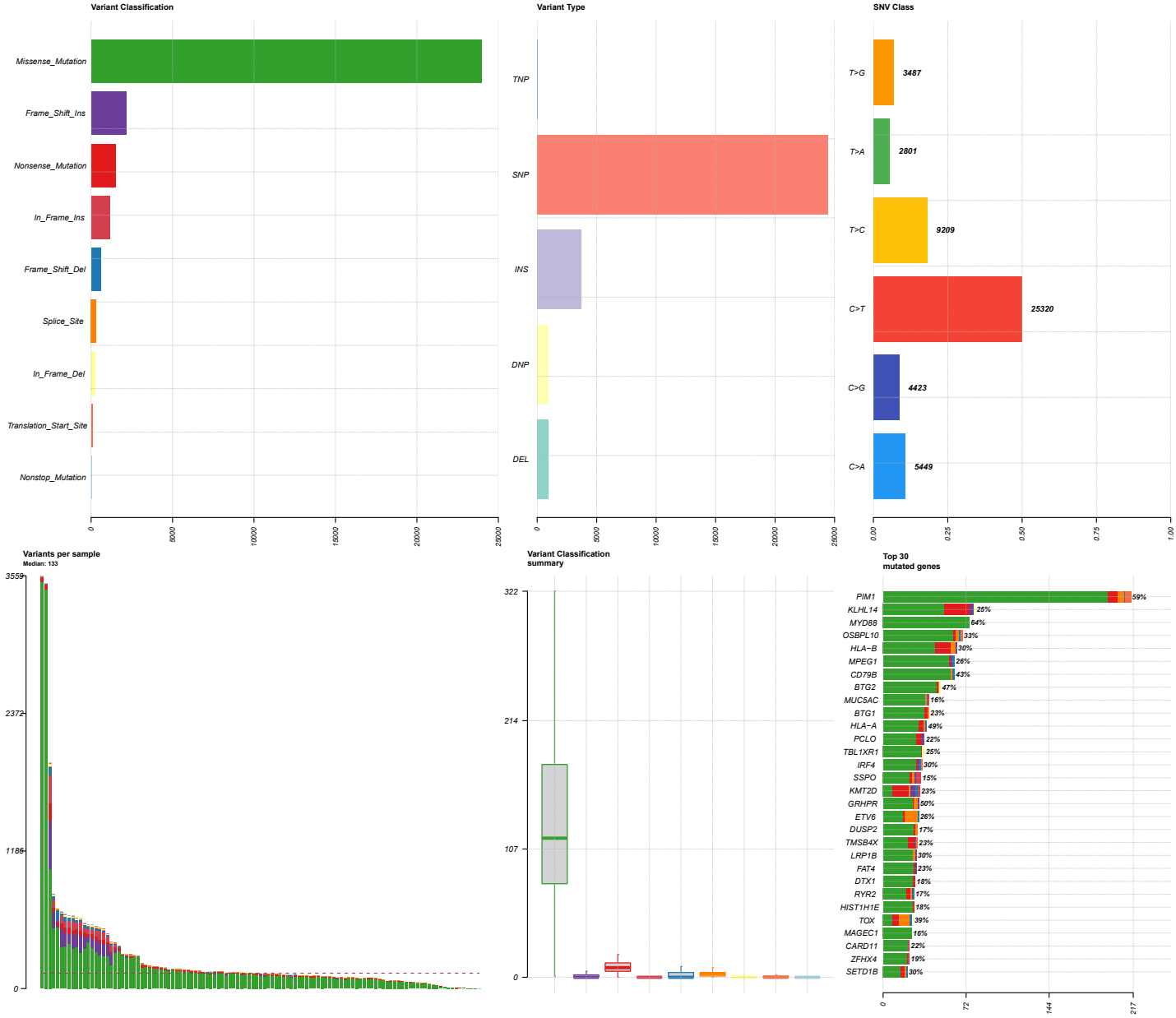

Variant Classification
Missense_Mutation
Frame_Shift_Ins
Nonsense_Mutation
In_Frame_Ins
Frame_Shift_Del
Splice_Site
In_Frame_Del
Translation_Start_Site
Nonstop_Mutation
0
5000
10000
15000
20000
25000
Variant Type
TNP
SNP
INS
DNP
DEL
0
5000
10000
15000
20000
25000
SNV Class
T>G
3487
T>A
2801
T>C
9209
C>T
25320
C>G
4423
C>A
5449
0.00
0.25
0.50
0.75
1.00
Variants per sample
Median: 133
3559
2372
1186
0
Variant Classification summary
322
214
107
0
Top 30 mutated genes
PIM1
59%
KLHL14
25%
MYD88
64%
OSBPL10
33%
HLA-B
30%
MPEG1
26%
CD79B
43%
BTG2
47%
MUC5AC
16%
BTG1
23%
HLA-A
49%
PCLO
22%
TBL1XR1
25%
IRF4
30%
SSPO
15%
KMT2D
23%
GRHPR
50%
ETV6
26%
DUSP2
17%
TMSB4X
23%
LRP1B
30%
FAT4
23%
DTX1
18%
RYR2
17%
HIST1H1E
18%
TOX
39%
MAGEC1
16%
CARD11
22%
ZFHX4
19%
SETD1B
30%
0
72
144
217

Figure S14

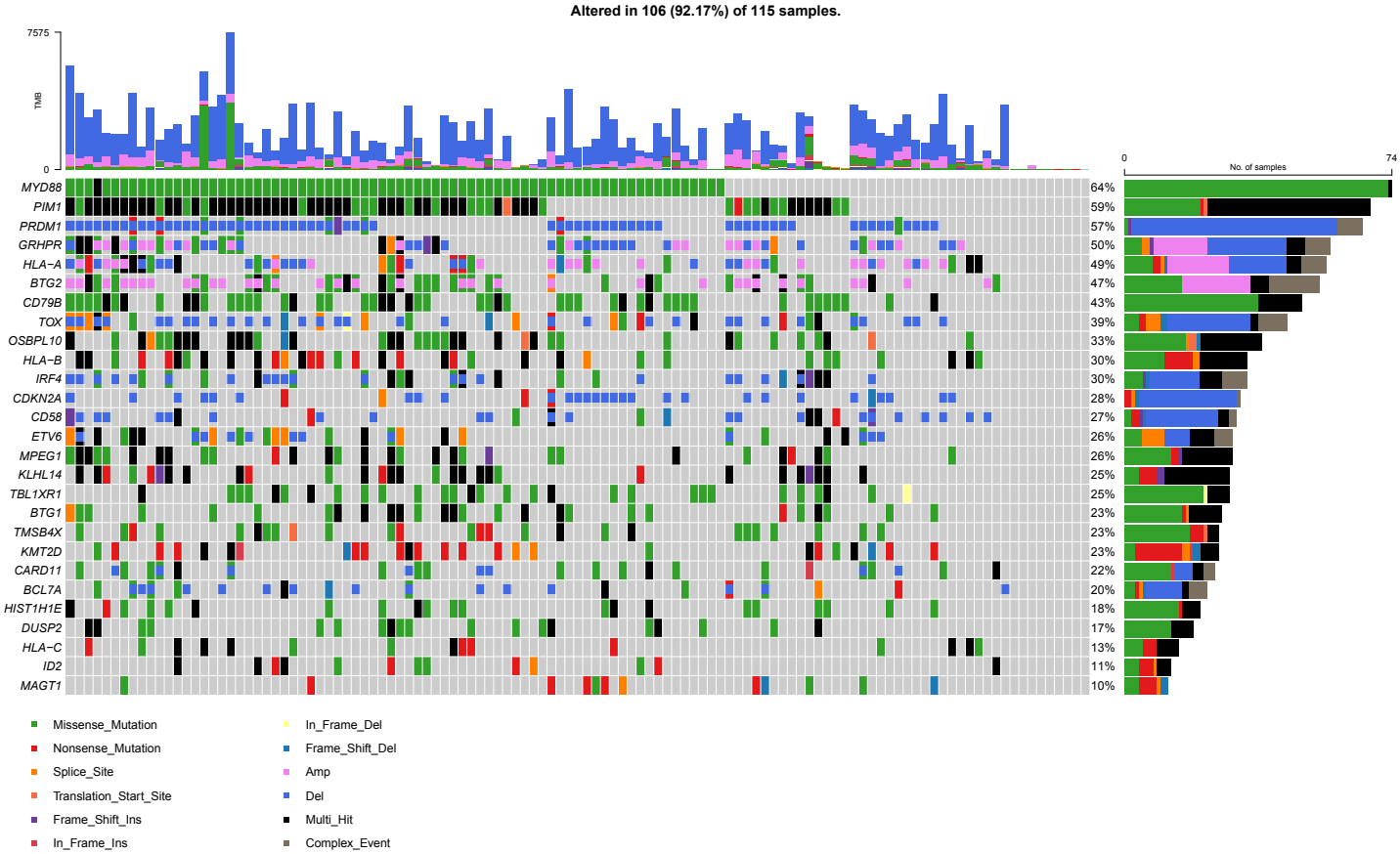

Altered in 106 (92.17%) of 115 samples.
7575
TMB
0
0
74
No. of samples
MYD88 64%
PIM1 59%
PRDM1 57%
GRHPR 50%
HLA−A 49%
BTG2 47%
CD79B 43%
TOX 39%
OSBPL10 33%
HLA−B 30%
IRF4 30%
CDKN2A 28%
CD58 27%
ETV6 26%
MPEG1 26%
KLHL14 25%
TBL1XR1 25%
BTG1 23%
TMSB4X 23%
KMT2D 23%
CARD11 22%
BCL7A 20%
HIST1H1E 18%
DUSP2 17%
HLA−C 13%
ID2 11%
MAGT1 10%
Missense_Mutation
Nonsense_Mutation
Splice_Site
Translation_Start_Site
Frame_Shift_Ins
In_Frame_Ins
In_Frame_Del
Frame_Shift_Del
Amp
Del
Multi_Hit
Complex_Event

Figure S15

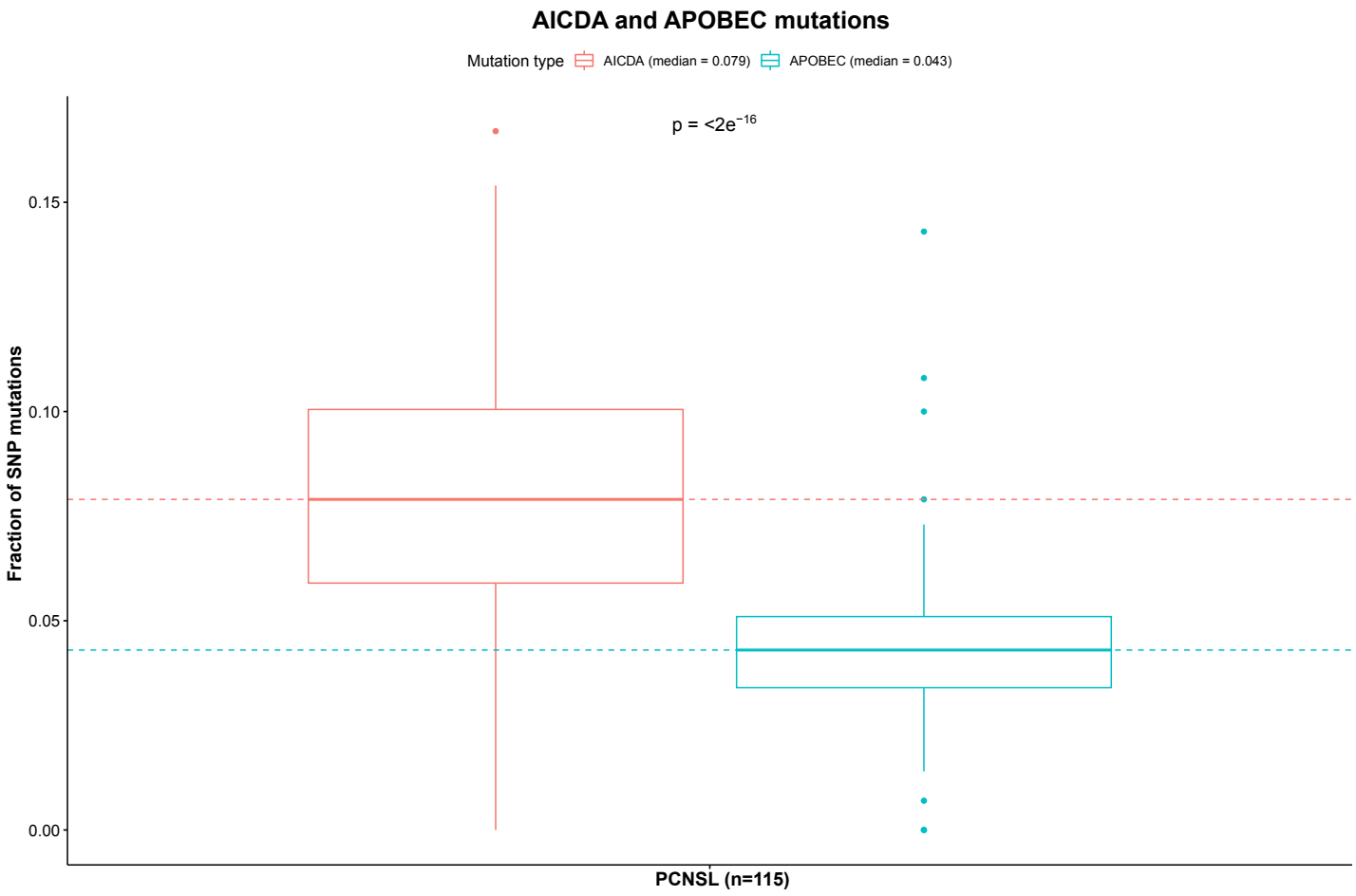

AICDA and APOBEC mutations
Mutation type
AICDA (median = 0.079)
APOBEC (median = 0.043)
p = <2e−16
0.15
0.10
0.05
0.00
Fraction of SNP mutations
PCNSL (n=115)

Figure S16

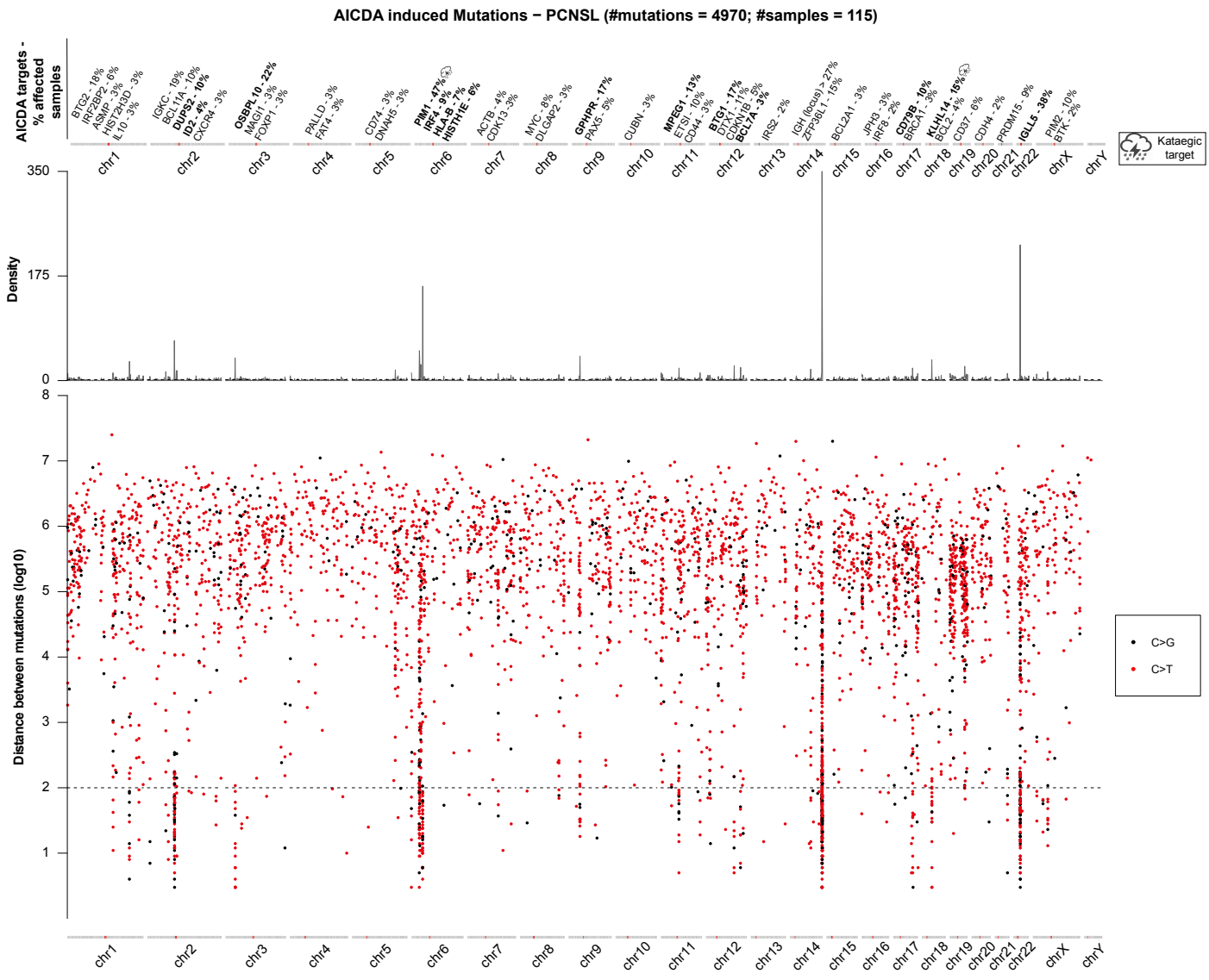

AICDA induced Mutations − PCNSL (#mutations = 4970; #samples = 115)
AICDA targets - % affected samples
BTG2 - 18%
IRF2BP2 - 6%
ASMP - 3%
HIST2H2BD - 3%
IL10 - 3%
IGKC - 19%
BCL11A - 10%
DUPS2 - 8%
ID2 - 6%
CXCR4 - 3%
OSBPL10 - 22%
NAGI1 - 3%
FOXP1 - 3%
PALLD - 3%
PAT4 - 3%
CD74 - 3%
DNAH5 - 3%
PIM1 - 47%
IRF4 - 9%
HLA-B - 7%
HIST1H - 6%
ACTB - 4%
CDK13 - 3%
MYC - 8%
DLGAP2 - 3%
GPHPR - 17%
PAX5 - 5%
CUBN - 3%
MPEG1 - 13%
ETS1 - 10%
CD44 - 3%
BTG1 - 17%
DTX1 - 17%
CDKN1B - 6%
BCL7A - 3%
IRS2 - 2%
IGH (locus) > 27%
ZFP36L1 - 15%
BCL2A1 - 3%
JPH3 - 3%
IRF8 - 2%
CD79B - 10%
BRCA1 - 3%
KLHL14 - 15%
BCL2 - 4%
CD37 - 6%
CDH4 - 2%
PRDM15 - 9%
IGLL5 - 38%
PIM2 - 10%
BTK - 2%
Kataegic target
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
Density
350
175
0
Distance between mutations (log10)
8
7
6
5
4
3
2
1
C>G
C>T

Figure S17

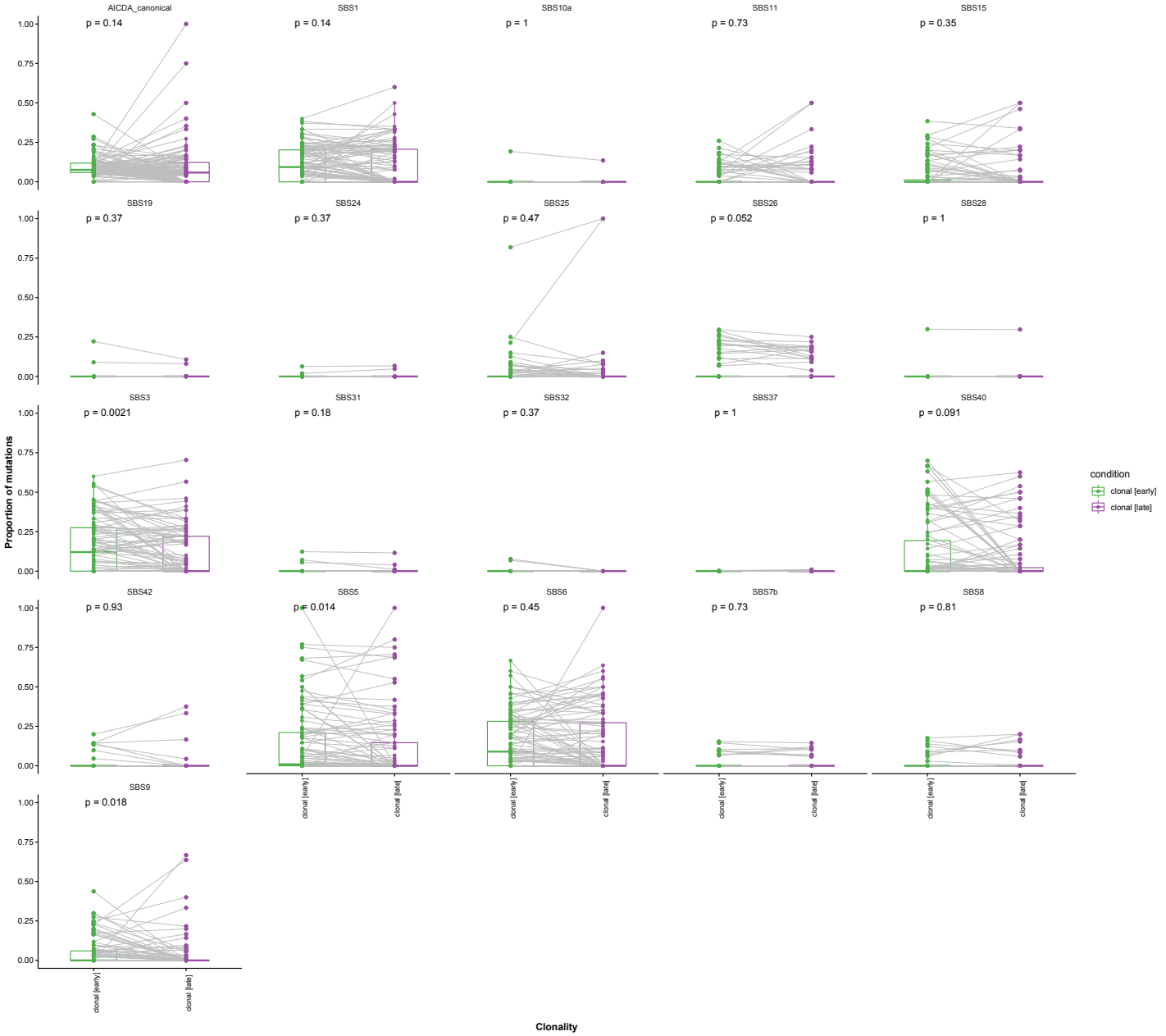

AICDA_canonical
p = 0.14
SBS1
p = 0.14
SBS10a
p = 1
SBS11
p = 0.73
SBS15
p = 0.35
SBS19
p = 0.37
SBS24
p = 0.37
SBS25
p = 0.47
SBS26
p = 0.052
SBS28
p = 1
SBS3
p = 0.0021
SBS31
p = 0.18
SBS32
p = 0.37
SBS37
p = 1
SBS40
p = 0.091
SBS42
p = 0.93
SBS5
p = 0.014
SBS6
p = 0.45
SBS7b
p = 0.73
SBS8
p = 0.81
SBS9
p = 0.018
1.00
0.75
0.50
0.25
0.00
clonal [early]
clonal [late]
Proportion of mutations
Clonality
condition
clonal [early]
clonal [late]

Figure S18

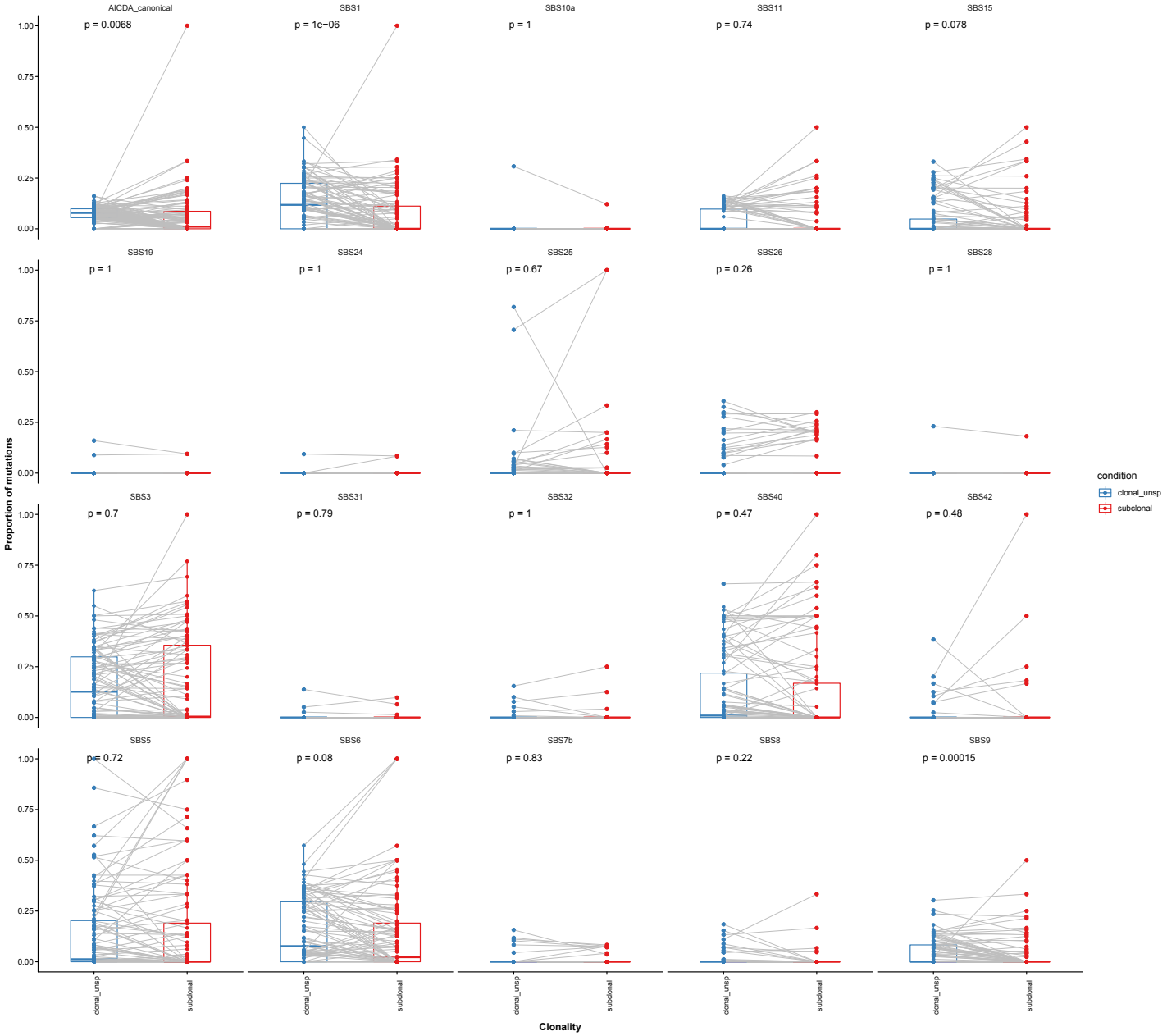

AICDA_canonical
p = 0.0068
SBS1
p = 1e−06
SBS10a
p = 1
SBS11
p = 0.74
SBS15
p = 0.078
SBS19
p = 1
SBS24
p = 1
SBS25
p = 0.67
SBS26
p = 0.26
SBS28
p = 1
SBS3
p = 0.7
SBS31
p = 0.79
SBS32
p = 1
SBS40
p = 0.47
SBS42
p = 0.48
SBS5
p = 0.72
SBS6
p = 0.08
SBS7b
p = 0.83
SBS8
p = 0.22
SBS9
p = 0.00015
Proportion of mutations
1.00
0.75
0.50
0.25
0.00
clonal_unsp
subclonal
Clonality
condition
clonal_unsp
subclonal

Figure S19

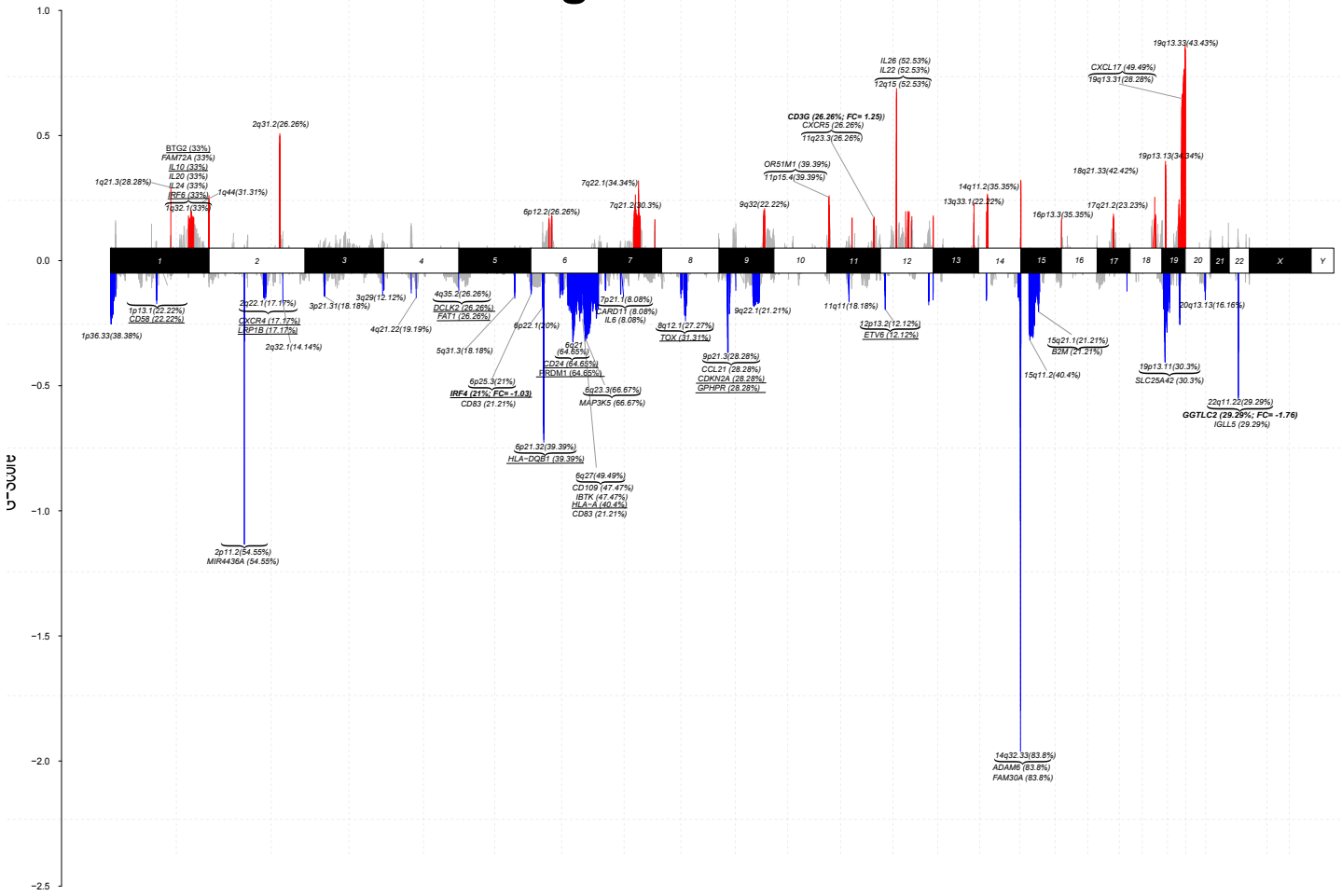

1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
1p36.33(38.38%)
1p13.1(22.22%)
CD58 (22.22%)
1q21.3(28.28%)
BTG2 (33%)
FAM72A (33%)
IL10 (33%)
IL20 (33%)
IL24 (33%)
IRF6 (33%)
1q32.1(33%)
1q44(31.31%)
2p11.2(54.55%)
MIR4436A (54.55%)
2q22.1(17.17%)
CXCR4 (17.17%)
LRP1B (17.17%)
2q32.1(14.14%)
2q31.2(26.26%)
3p21.31(18.18%)
3q29(12.12%)
4q21.22(19.19%)
4q35.2(26.26%)
DCLK2 (26.26%)
FAT1 (26.26%)
5q31.3(18.18%)
6p25.3(21%)
IRF4 (21%; FC= -1.03)
CD83 (21.21%)
6p22.1(20%)
6p12.2(26.26%)
6p21.32(39.39%)
HLA-DQB1 (39.39%)
6q21 (64.65%)
CD24 (64.65%)
PRDM1 (64.65%)
6q23.3(66.67%)
MAP3K5 (66.67%)
6q27(49.49%)
CD109 (47.47%)
IBTK (47.47%)
HLA-A (40.4%)
CD83 (21.21%)
7p22.1(34.34%)
7q21.2(30.3%)
7p21.1(8.08%)
CARD11 (8.08%)
IL6 (8.08%)
8q12.1(27.27%)
TOX (31.31%)
9p21.3(28.28%)
CCL21 (28.28%)
CDKN2A (28.28%)
GPHPR (28.28%)
9q32(22.22%)
9q22.1(21.21%)
CD3G (26.26%; FC= 1.25))
CXCR5 (26.26%)
11q23.3(26.26%)
OR51M1 (39.39%)
11p15.4(39.39%)
11q11(18.18%)
IL26 (52.53%)
IL22 (52.53%)
12q15 (52.53%)
12p13.2(12.12%)
ETV6 (12.12%)
13q33.1(22.22%)
14q11.2(35.35%)
14q32.33(83.8%)
ADAM6 (83.8%)
FAM30A (83.8%)
15q11.2(40.4%)
15q21.1(21.21%)
B2M (21.21%)
16p13.3(35.35%)
17q21.2(23.23%)
18q21.33(42.42%)
19q13.33(43.43%)
CXCL17 (49.49%)
19q13.31(28.28%)
19p13.13(34.34%)
19p13.11(30.3%)
SLC25A42 (30.3%)
20q13.13(16.16%)
22q11.22(29.29%)
GGTLC2 (29.29%; FC= -1.76)
IGLL5 (29.29%)

Figure S20

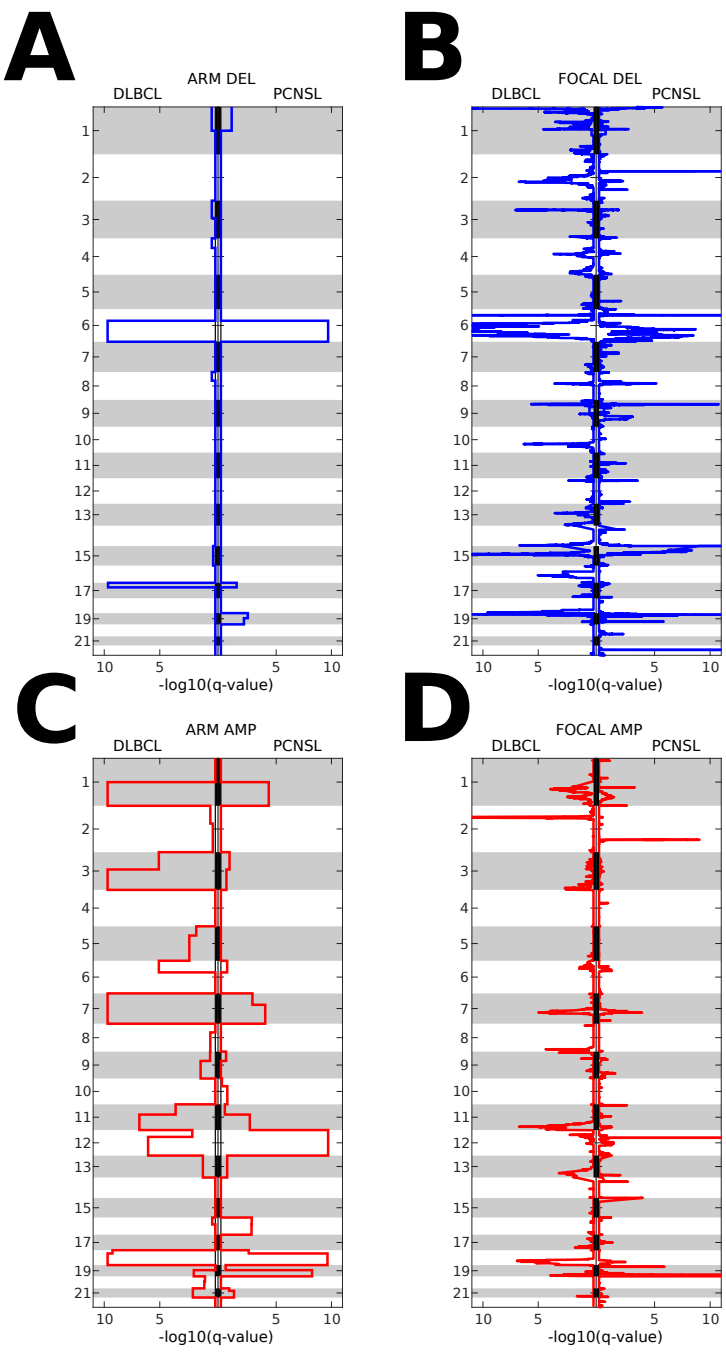

A
ARM DEL
DLBCL
PCNSL
B
FOCAL DEL
DLBCL
PCNSL
C
ARM AMP
DLBCL
PCNSL
D
FOCAL AMP
DLBCL
PCNSL
-log10(q-value)

Figure S21

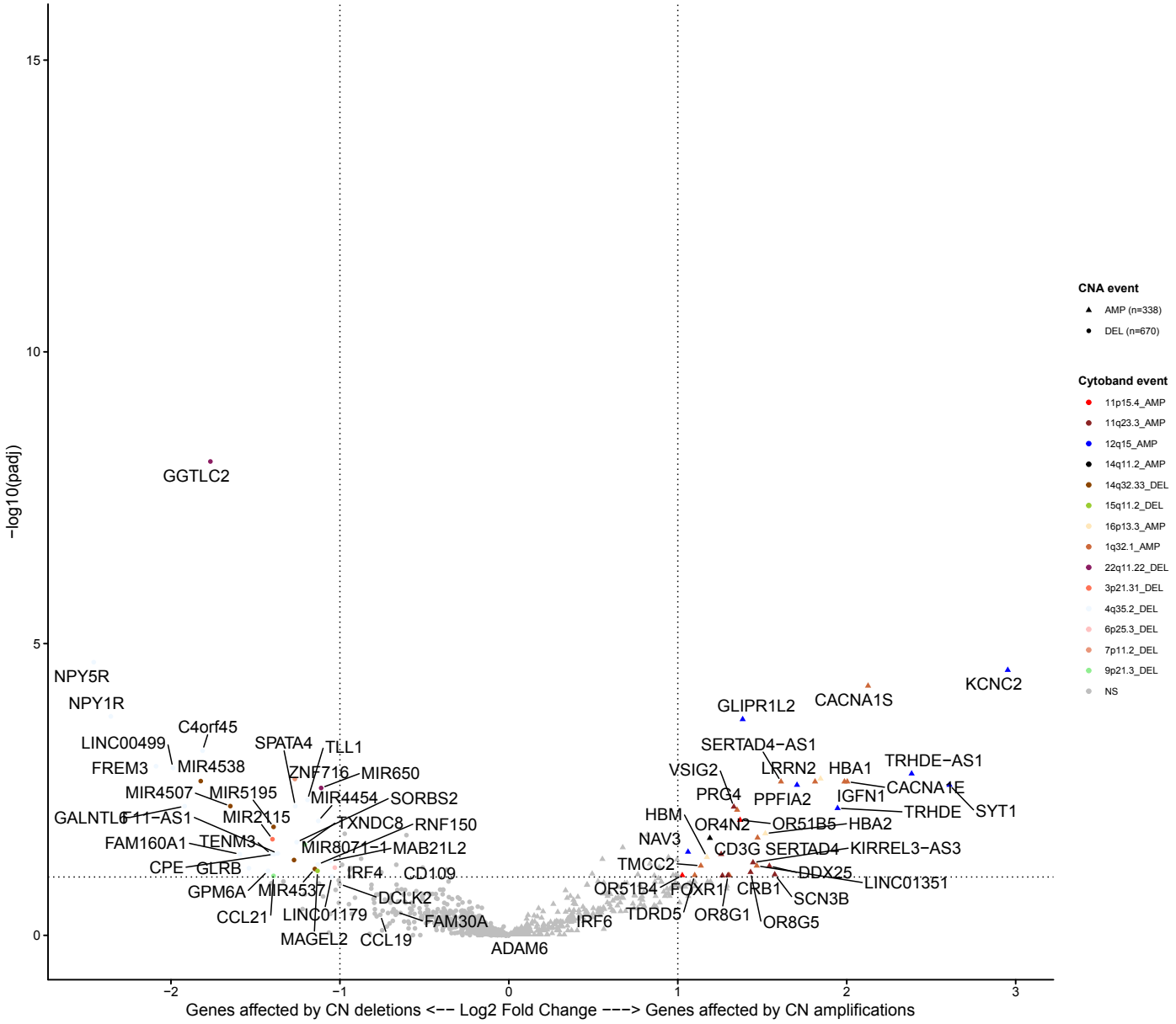

−log10(padj)
15
10
5
0
−2
−1
0
1
2
3
Genes affected by CN deletions <−− Log2 Fold Change −−−> Genes affected by CN amplifications
GGTLC2
NPY5R
NPY1R
C4orf45
LINC00499
SPATA4
TLL1
FREM3
MIR4538
ZNF716
MIR650
MIR4507
MIR5195
MIR4454
SORBS2
GALNTLG
F11−AS1
MIR2115
TXNDC8
RNF150
FAM160A1
TENM3
MIR8071−1
MAB21L2
CPE
GLRB
IRF4
CD109
GPM6A
MIR4537
DCLK2
CCL21
LINC01179
FAM30A
MAGEL2
CCL19
ADAM6
IRF6
TDRD5
OR51B4
FOXR1
CRB1
OR8G1
OR8G5
SCN3B
TMCC2
NAV3
HBM
OR4N2
OR51B5
HBA2
CD3G
SERTAD4
KIRREL3−AS3
DDX25
LINC01351
VSIG2
PRG4
PPFIA2
SERTAD4−AS1
LRRN2
HBA1
TRHDE−AS1
IGFN1
CACNA1E
TRHDE
SYT1
GLIPR1L2
CACNA1S
KCNC2
CNA event
AMP (n=338)
DEL (n=670)
Cytoband event
11p15.4_AMP
11q23.3_AMP
12q15_AMP
14q11.2_AMP
14q32.33_DEL
15q11.2_DEL
16p13.3_AMP
1q32.1_AMP
22q11.22_DEL
3p21.31_DEL
4q35.2_DEL
6p25.3_DEL
7p11.2_DEL
9p21.3_DEL
NS

Figure S22

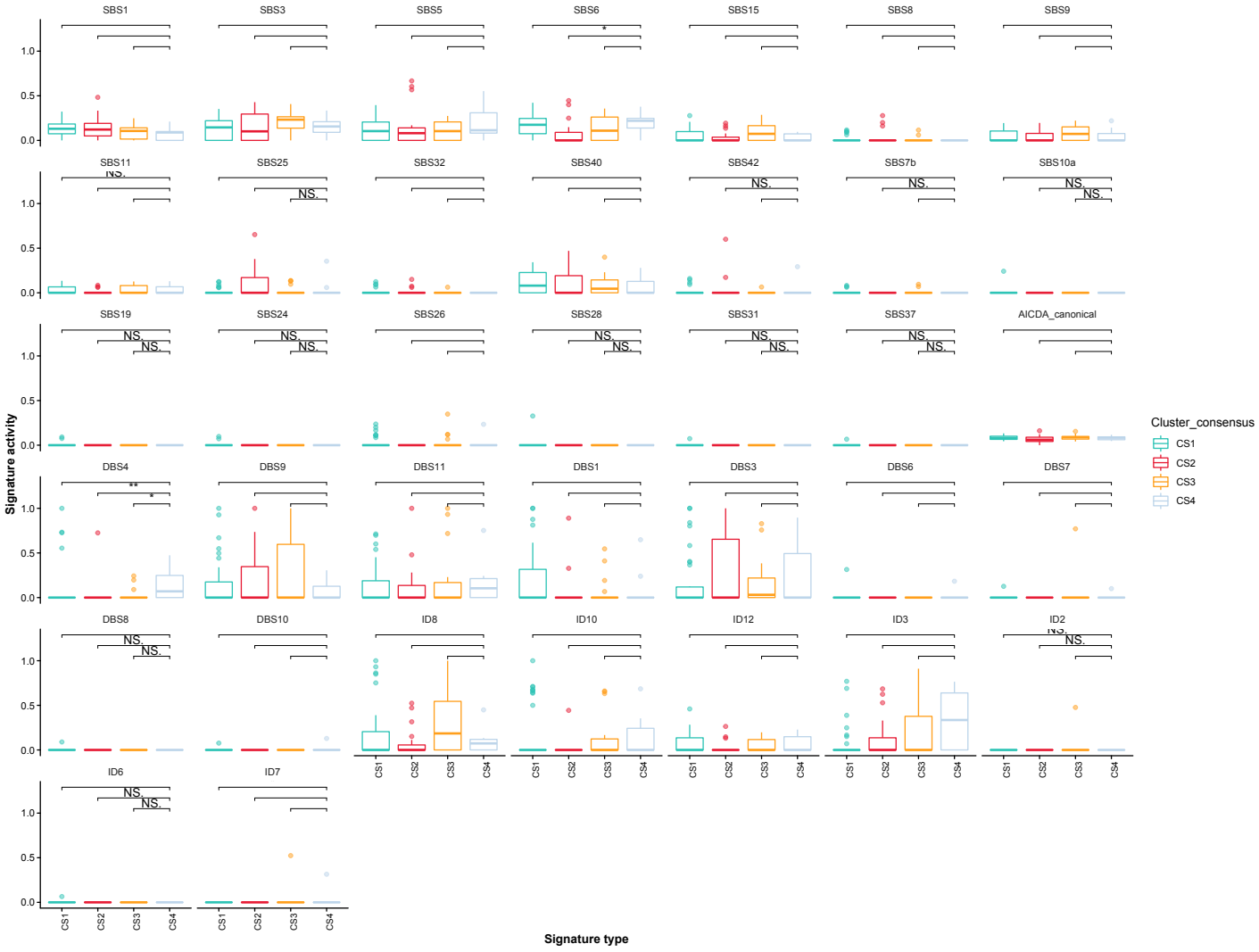

SBS1
SBS3
SBS5
SBS6
SBS15
SBS8
SBS9
SBS11
SBS25
SBS32
SBS40
SBS42
SBS7b
SBS10a
SBS19
SBS24
SBS26
SBS28
SBS31
SBS37
AICDA_canonical
DBS4
DBS9
DBS11
DBS1
DBS3
DBS6
DBS7
DBS8
DBS10
ID8
ID10
ID12
ID3
ID2
ID6
ID7
NS.
*
**
CS1
CS2
CS3
CS4
Signature activity
Signature type
Cluster_consensus

Figure S23

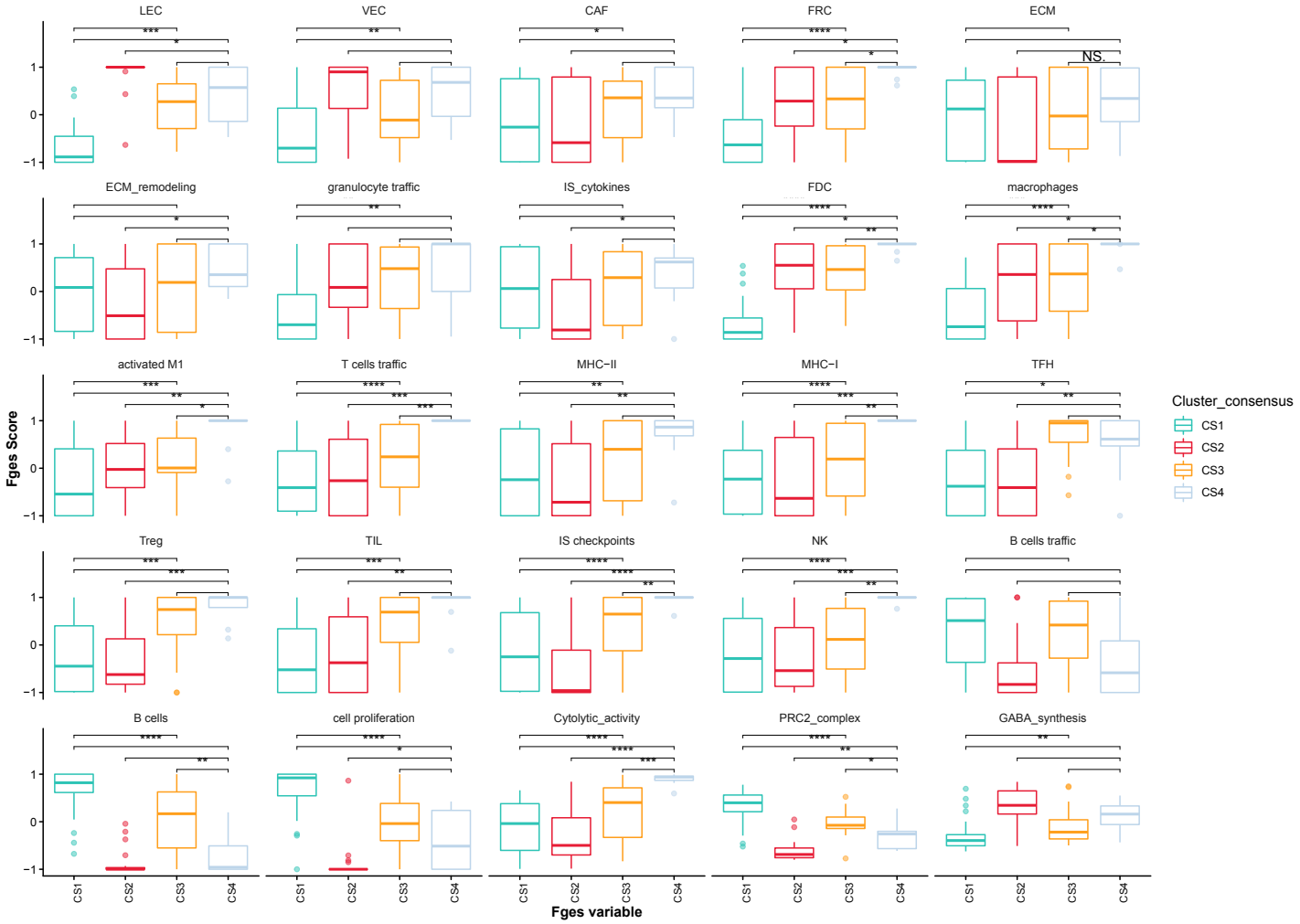

LEC
VEC
CAF
FRC
ECM
NS.
ECM_remodeling
granulocyte traffic
IS_cytokines
FDC
macrophages
activated M1
T cells traffic
MHC−II
MHC−I
TFH
Treg
TIL
IS checkpoints
NK
B cells traffic
B cells
cell proliferation
Cytolytic_activity
PRC2_complex
GABA_synthesis
Fges Score
Fges variable
CS1
CS2
CS3
CS4
Cluster_consensus

Figure S24

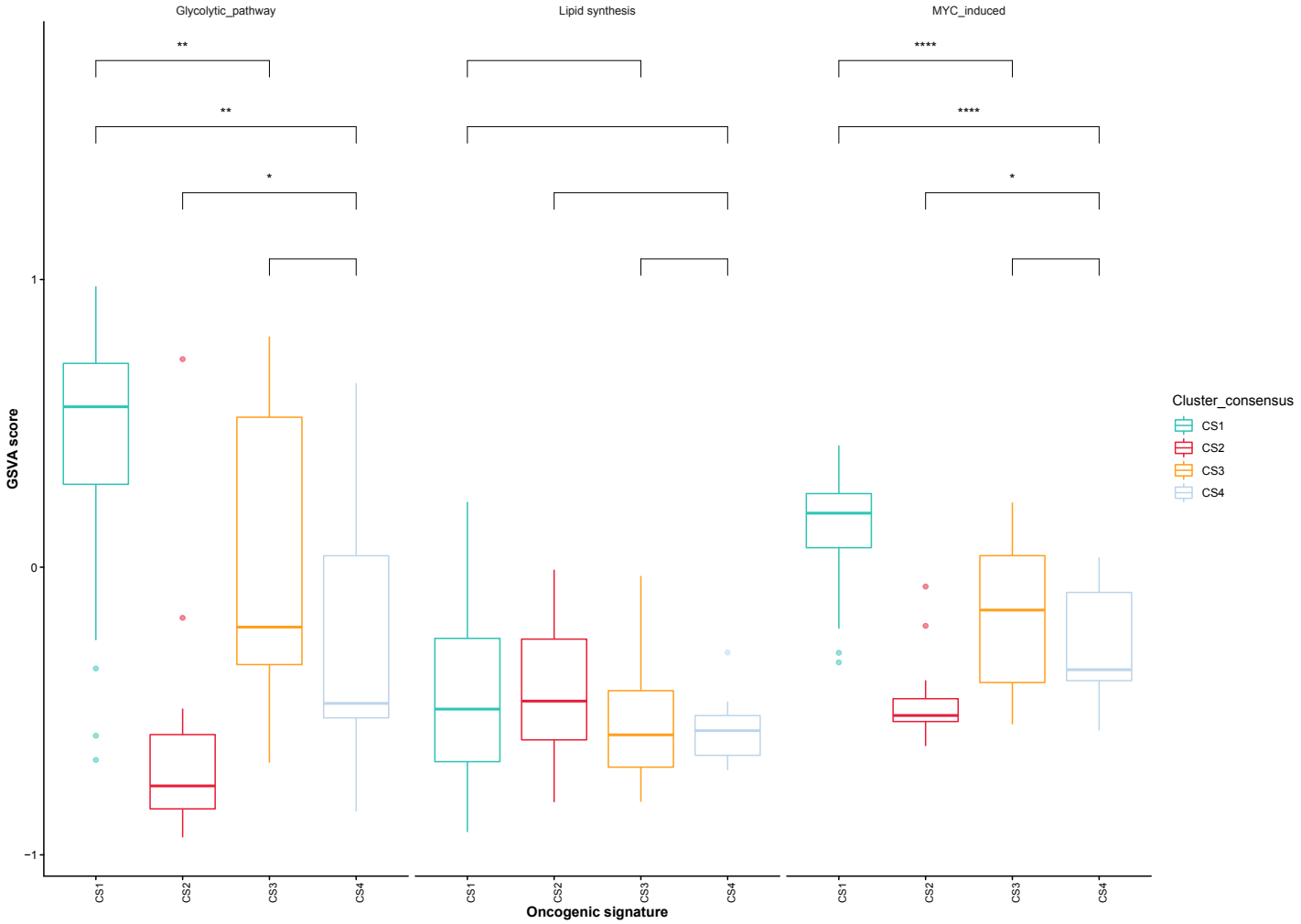

Glycolytic_pathway
Lipid synthesis
MYC_induced
**
**
*
****
****
*
GSVA score
1
0
−1
CS1
CS2
CS3
CS4
Oncogenic signature
Cluster_consensus
CS1
CS2
CS3
CS4

Figure S25


Androgen
EGFR
Estrogen
Hypoxia
JAK−STAT
MAPK
NFkB
p53
PI3K
TGFb
TNFa
Trail
VEGF
WNT
Progeny Score
Progeny variable
CS1
CS2
CS3
CS4
Cluster_consensus

Figure S26

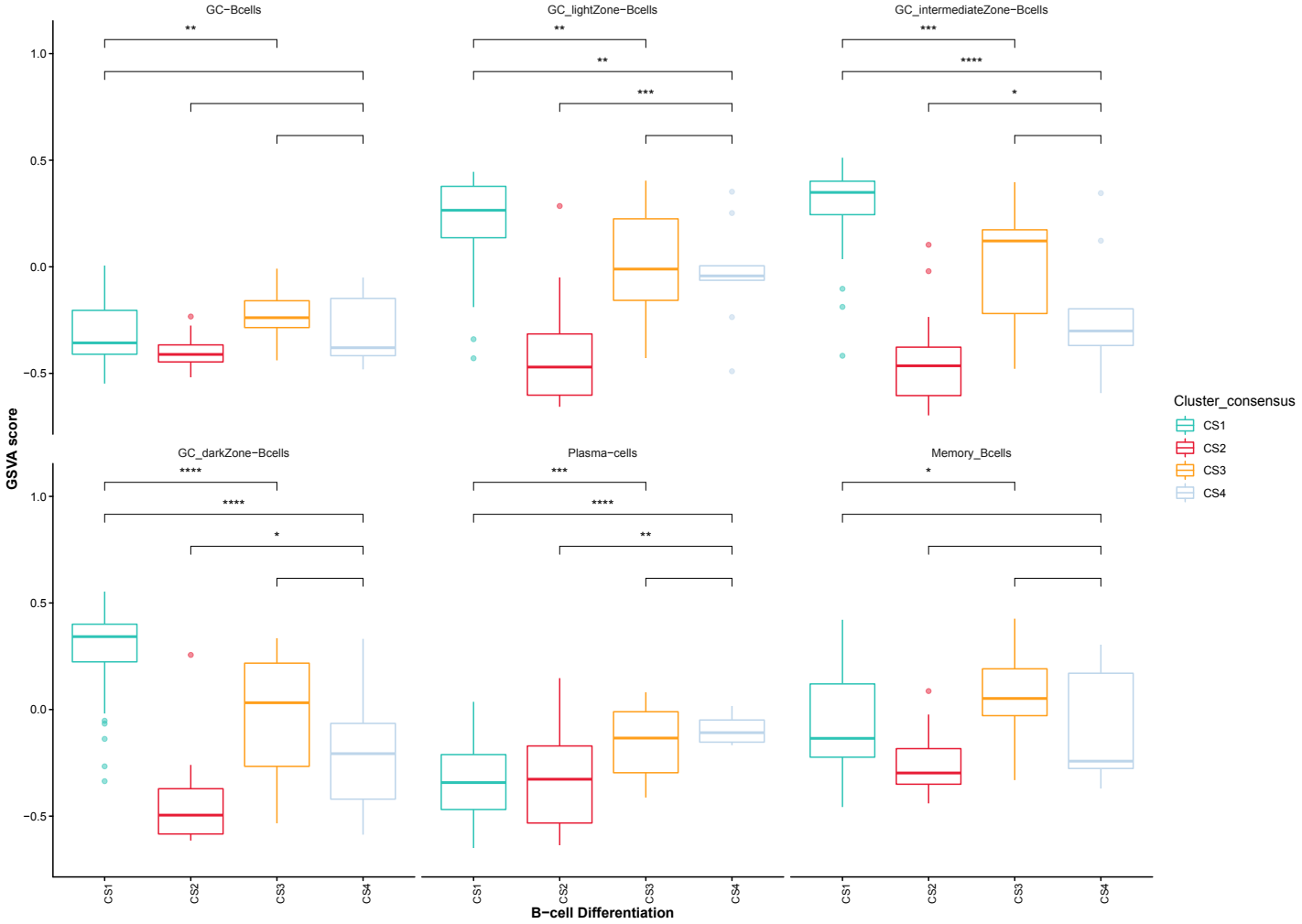

GC−Bcells
GC_lightZone−Bcells
GC_intermediateZone−Bcells
GC_darkZone−Bcells
Plasma−cells
Memory_Bcells
GSVA score
B−cell Differentiation
CS1
CS2
CS3
CS4
Cluster_consensus

Figure S27

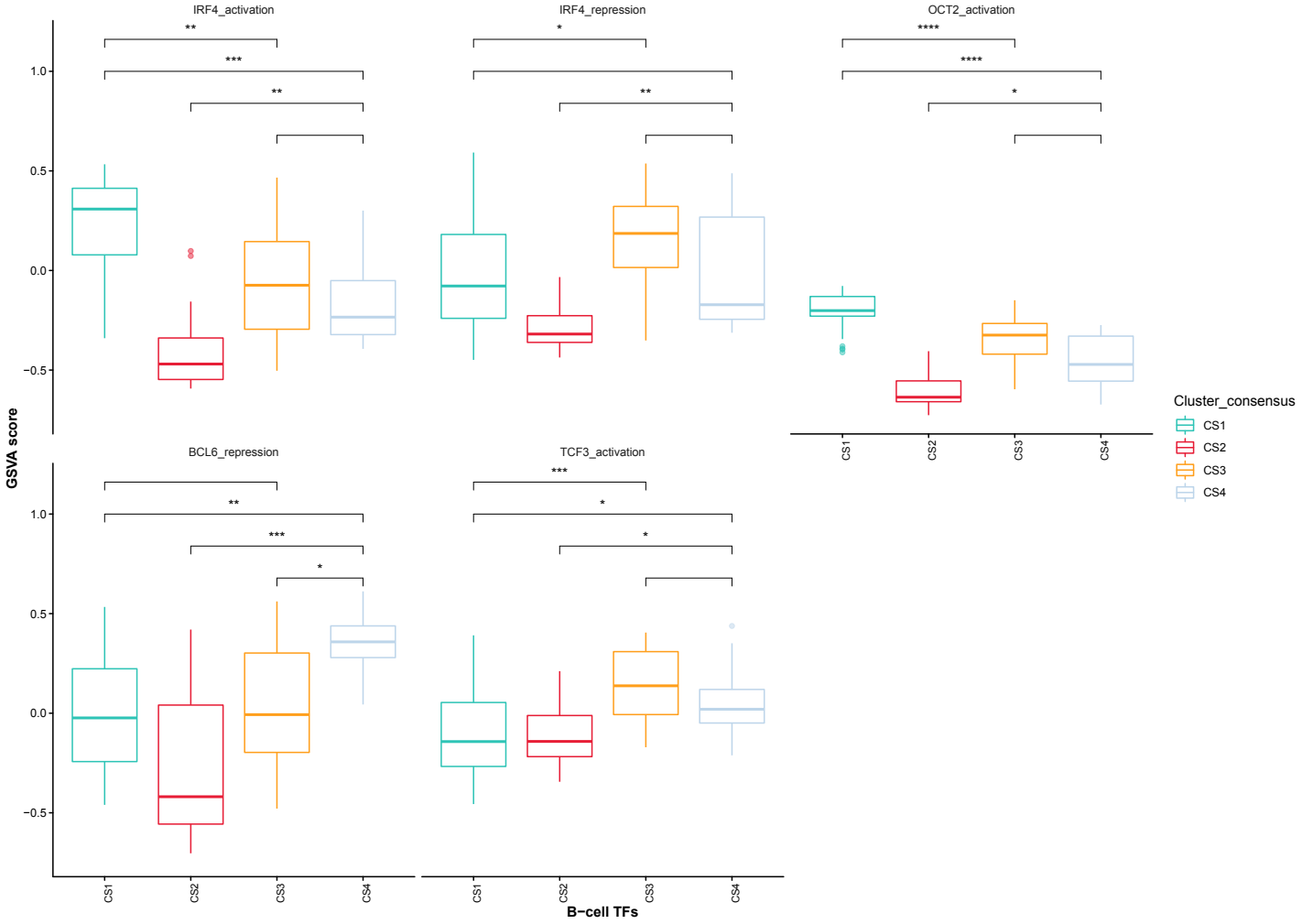

IRF4_activation
IRF4_repression
OCT2_activation
BCL6_repression
TCF3_activation
GSVA score
B−cell TFs
CS1
CS2
CS3
CS4
Cluster_consensus

Figure S28

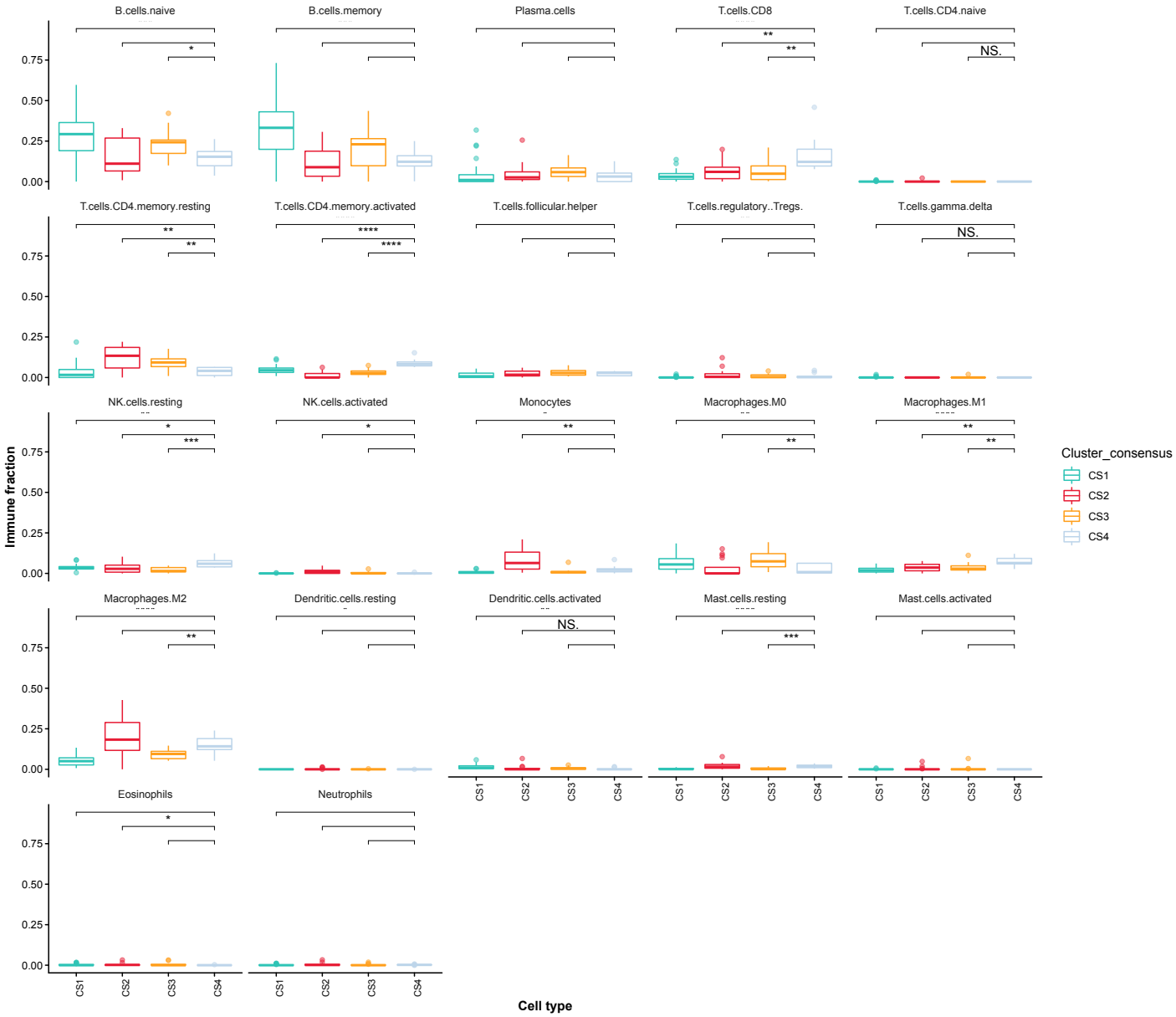

B.cells.naive
B.cells.memory
Plasma.cells
T.cells.CD8
T.cells.CD4.naive
T.cells.CD4.memory.resting
T.cells.CD4.memory.activated
T.cells.follicular.helper
T.cells.regulatory..Tregs.
T.cells.gamma.delta
NK.cells.resting
NK.cells.activated
Monocytes
Macrophages.M0
Macrophages.M1
Macrophages.M2
Dendritic.cells.resting
Dendritic.cells.activated
Mast.cells.resting
Mast.cells.activated
Eosinophils
Neutrophils
NS.
Immune fraction
0.00
0.25
0.50
0.75
CS1
CS2
CS3
CS4
Cell type
Cluster_consensus

Figure S29

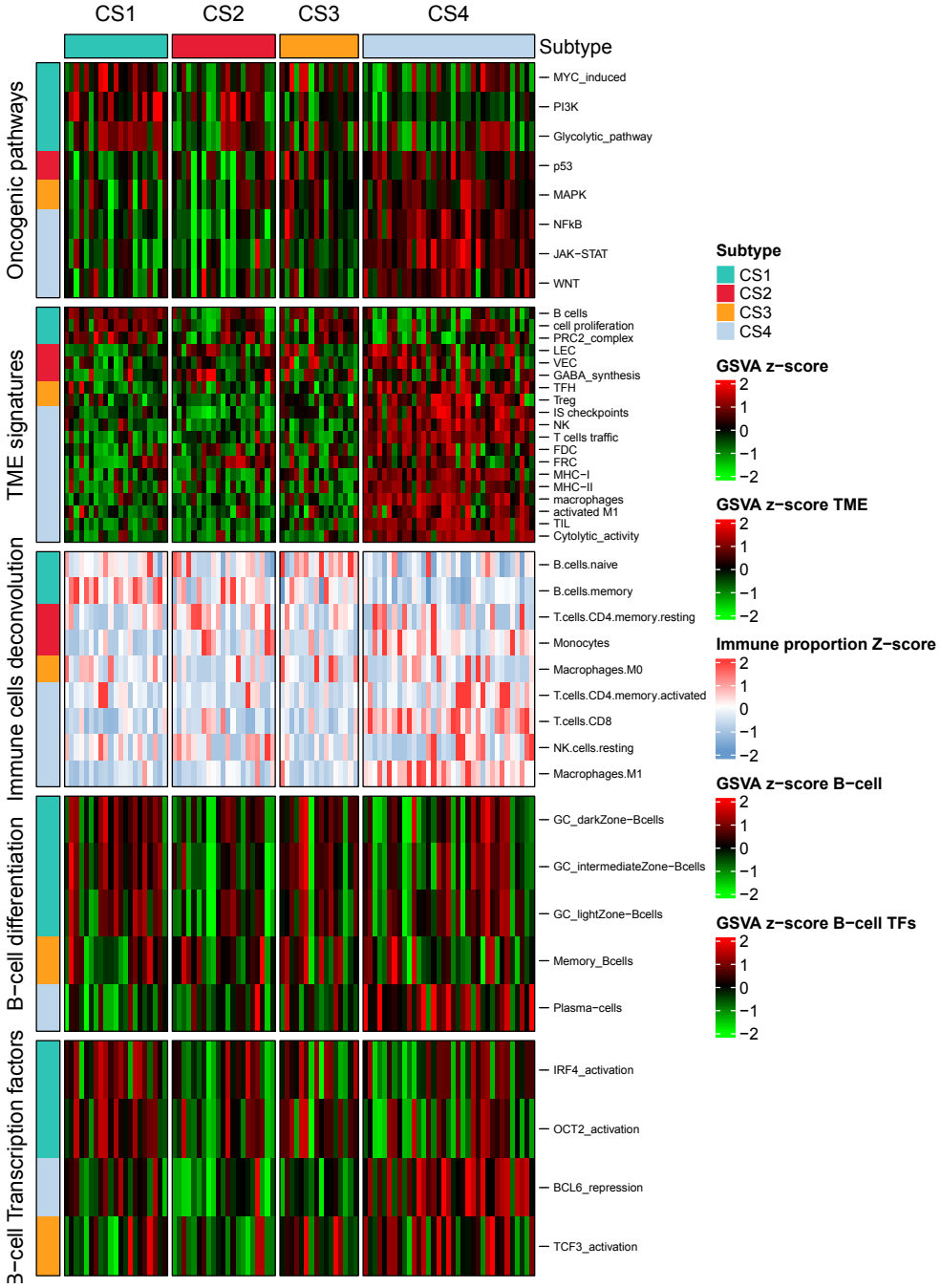

CS1
CS2
CS3
CS4
Subtype
Oncogenic pathways
MYC_induced
PI3K
Glycolytic_pathway
p53
MAPK
NFkB
JAK−STAT
WNT
TME signatures
B cells
cell proliferation
PRC2_complex
LEC
VEC
GABA_synthesis
TFH
Treg
IS checkpoints
NK
T cells traffic
FDC
FRC
MHC−I
MHC−II
macrophages
activated M1
TIL
Cytolytic_activity
Immune cells deconvolution
B.cells.naive
B.cells.memory
T.cells.CD4.memory.resting
Monocytes
Macrophages.M0
T.cells.CD4.memory.activated
T.cells.CD8
NK.cells.resting
Macrophages.M1
B−cell differentiation
GC_darkZone−Bcells
GC_intermediateZone−Bcells
GC_lightZone−Bcells
Memory_Bcells
Plasma−cells
3−cell Transcription factors
IRF4_activation
OCT2_activation
BCL6_repression
TCF3_activation
Subtype
CS1
CS2
CS3
CS4
GSVA z−score
2
1
0
−1
−2
GSVA z−score TME
2
1
0
−1
−2
Immune proportion Z−score
2
1
0
−1
−2
GSVA z−score B−cell
2
1
0
−1
−2
GSVA z−score B−cell TFs
2
1
0
−1
−2

Figure S30

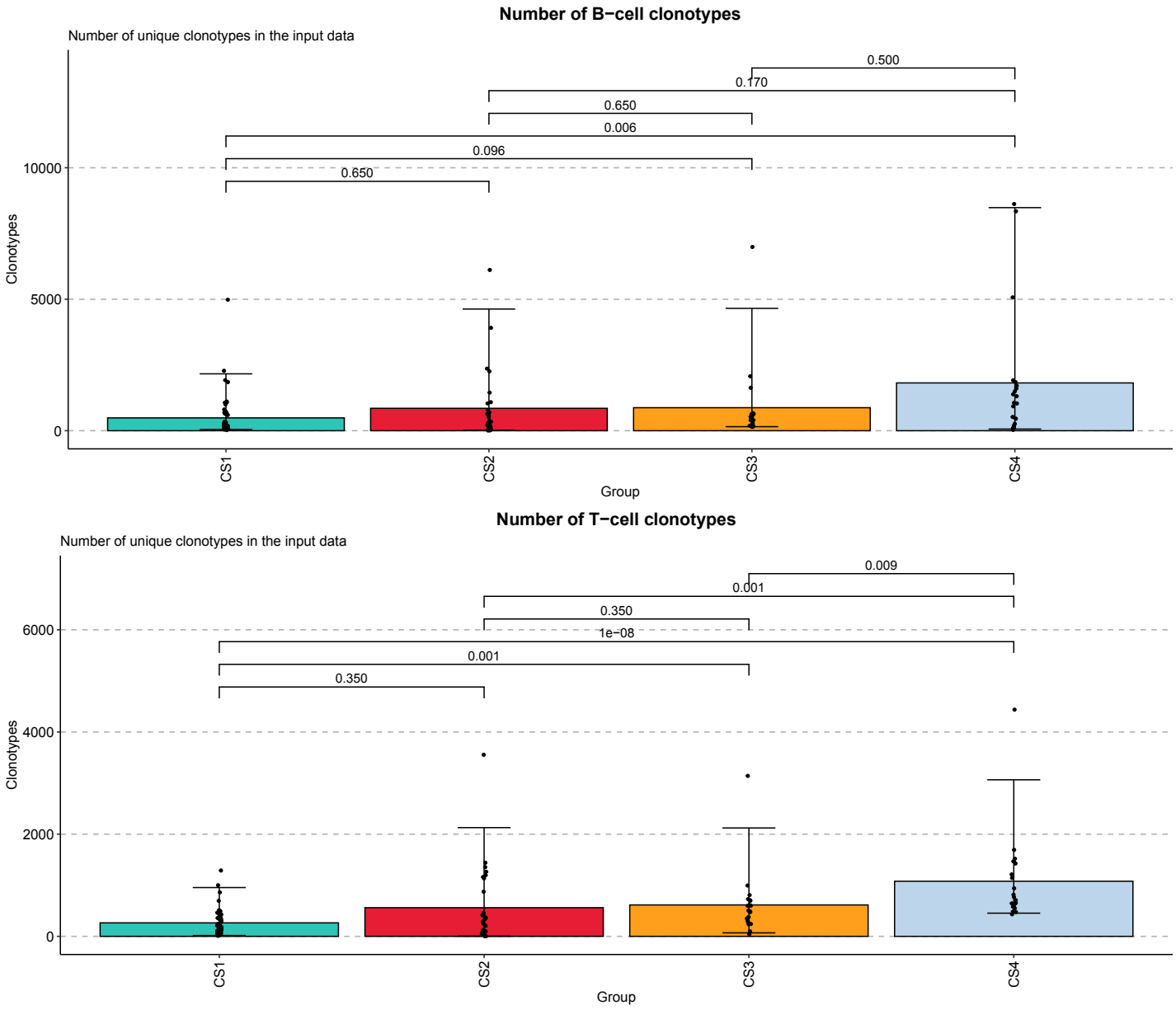

Number of B−cell clonotypes
Number of unique clonotypes in the input data
0.500
0.170
0.650
0.006
0.096
0.650
10000
5000
0
Clonotypes
CS1
CS2
CS3
CS4
Group
Number of T−cell clonotypes
Number of unique clonotypes in the input data
0.009
0.001
0.350
1e−08
0.001
0.350
6000
4000
2000
0
Clonotypes
CS1
CS2
CS3
CS4
Group

Figure S31

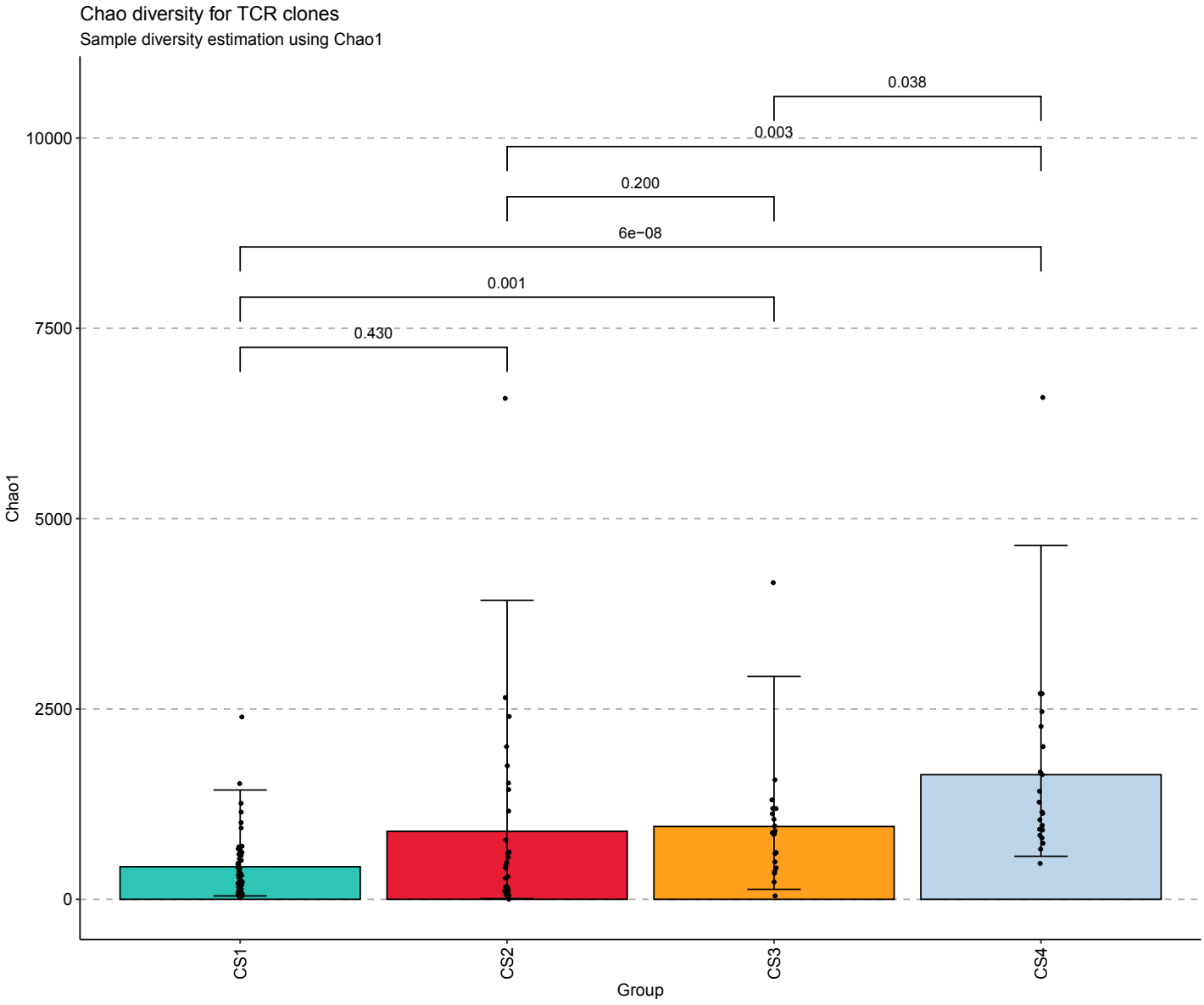

Chao diversity for TCR clones
Sample diversity estimation using Chao1
0.038
0.003
0.200
6e−08
0.001
0.430
10000
7500
5000
2500
0
Chao1
CS1
CS2
CS3
CS4
Group

Figure S32

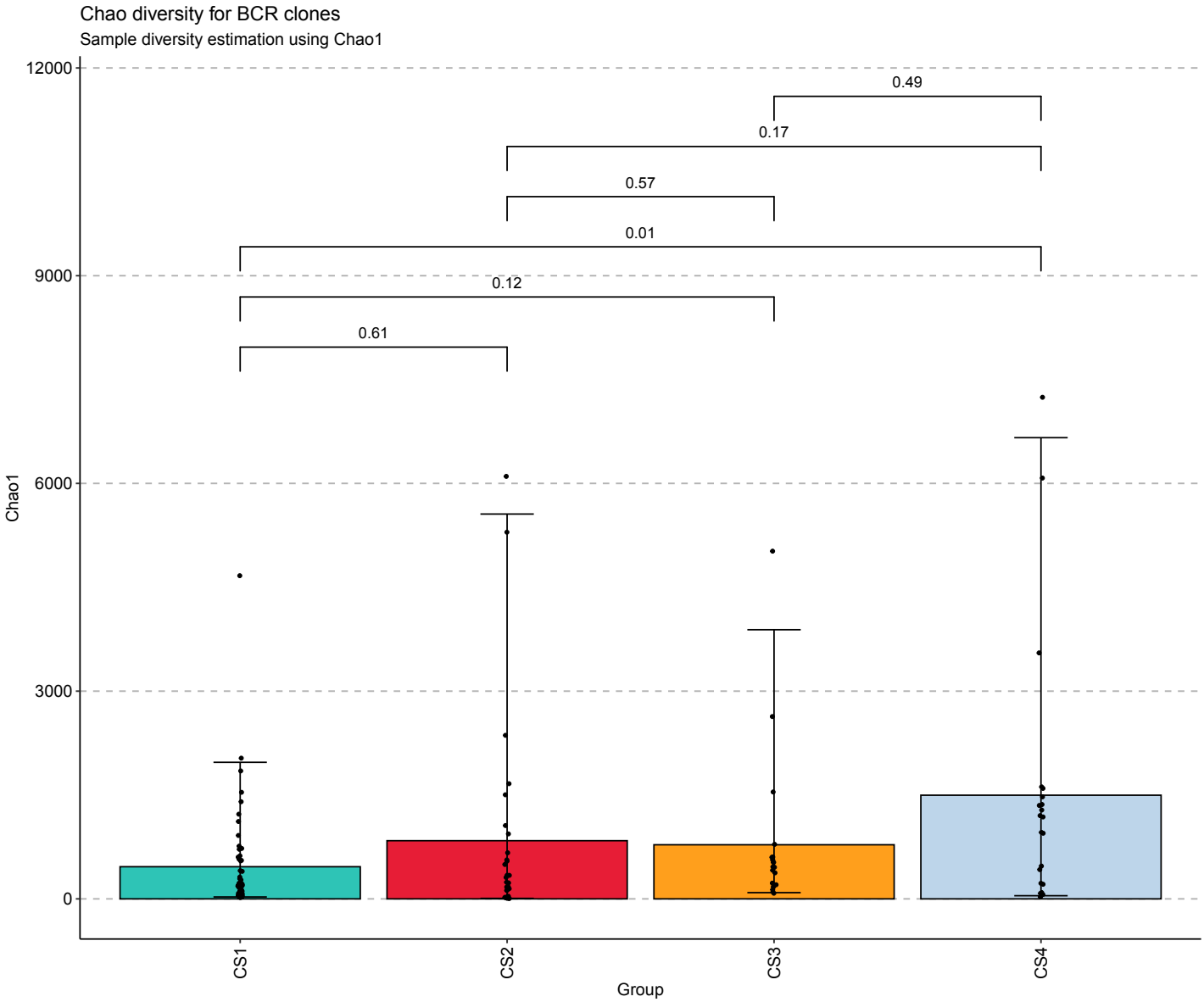

Chao diversity for BCR clones
Sample diversity estimation using Chao1
12000
9000
6000
3000
0
Chao1
0.49
0.17
0.57
0.01
0.12
0.61
CS1
CS2
CS3
CS4
Group

Figure S33

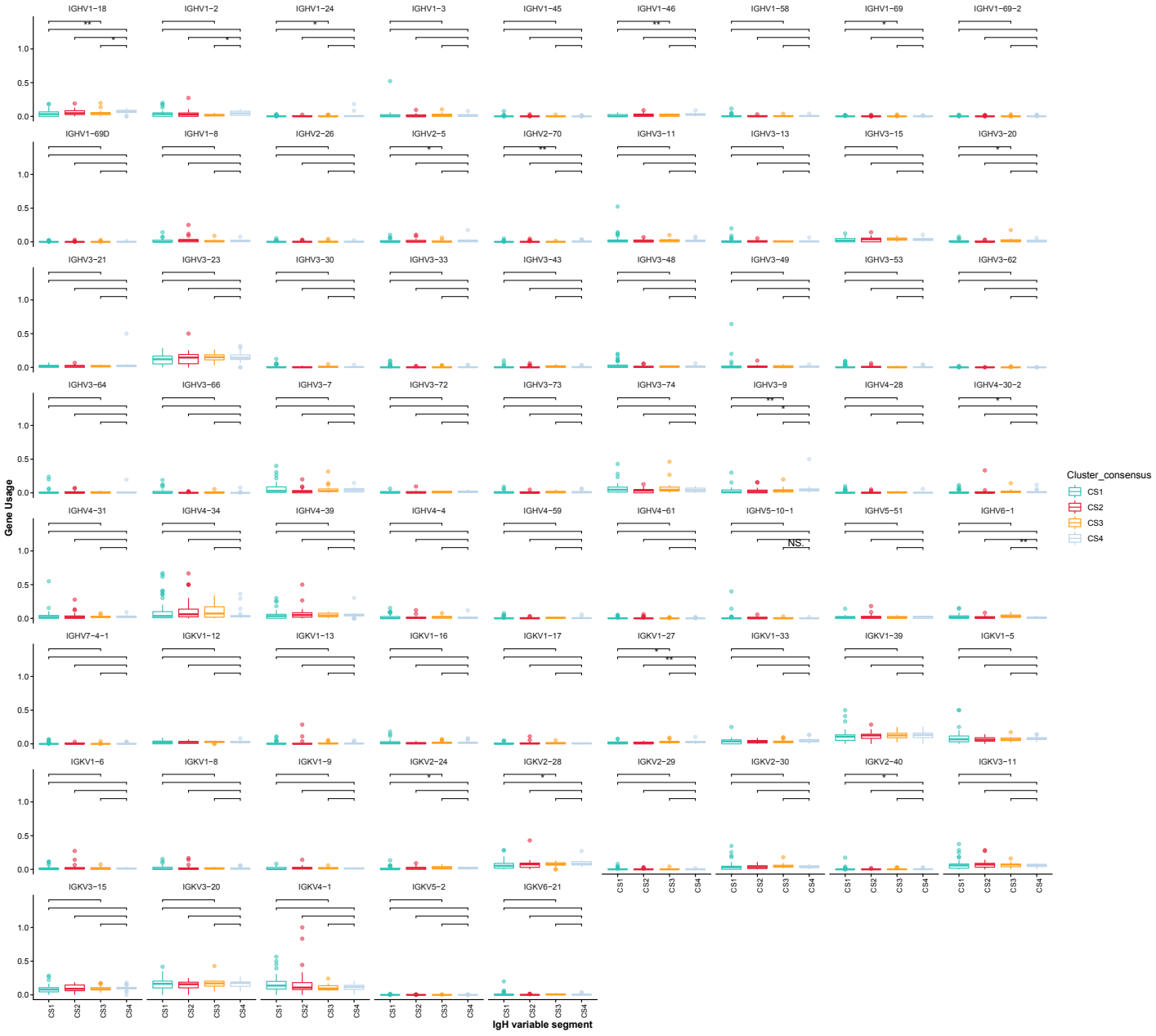

IGHV1−18
IGHV1−2
IGHV1−24
IGHV1−3
IGHV1−45
IGHV1−46
IGHV1−58
IGHV1−69
IGHV1−69−2
IGHV1−69D
IGHV1−8
IGHV2−26
IGHV2−5
IGHV2−70
IGHV3−11
IGHV3−13
IGHV3−15
IGHV3−20
IGHV3−21
IGHV3−23
IGHV3−30
IGHV3−33
IGHV3−43
IGHV3−48
IGHV3−49
IGHV3−53
IGHV3−62
IGHV3−64
IGHV3−66
IGHV3−7
IGHV3−72
IGHV3−73
IGHV3−74
IGHV3−9
IGHV4−28
IGHV4−30−2
IGHV4−31
IGHV4−34
IGHV4−39
IGHV4−4
IGHV4−59
IGHV4−61
IGHV5−10−1
IGHV5−51
IGHV6−1
IGHV7−4−1
IGKV1−12
IGKV1−13
IGKV1−16
IGKV1−17
IGKV1−27
IGKV1−33
IGKV1−39
IGKV1−5
IGKV1−6
IGKV1−8
IGKV1−9
IGKV2−24
IGKV2−28
IGKV2−29
IGKV2−30
IGKV2−40
IGKV3−11
IGKV3−15
IGKV3−20
IGKV4−1
IGKV5−2
IGKV6−21
NS.
Gene Usage
IgH variable segment
CS1
CS2
CS3
CS4
Cluster_consensus

Figure S34

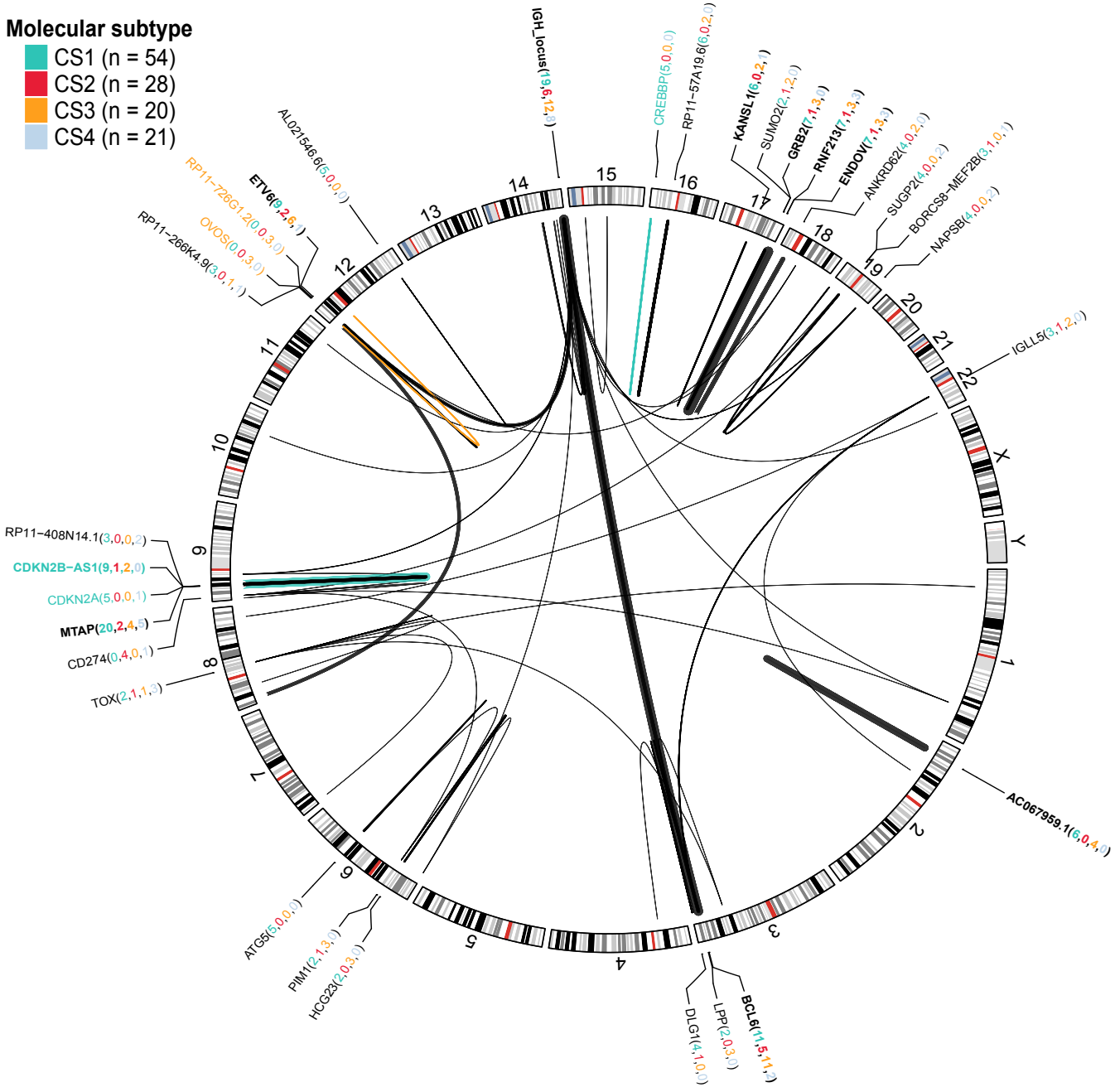

Molecular subtype
CS1 (n = 54)
CS2 (n = 28)
CS3 (n = 20)
CS4 (n = 21)
IGH_locus(19,6,12,1)
CREBBP(5,0,0,0)
RP11−57A19.6(6,0,2,1)
KANSL1(6,0,2,1)
SUMO2(2,1,2,1)
GRB2(7,1,3,1)
RNF213(7,1,3,1)
ENDOV(7,1,3,1)
ANKRD62(4,0,2,1)
SUGP2(4,0,0,2)
BORCS8−MEF2B(3,1,0,1)
NAPSB(4,0,0,2)
IGLL5(3,1,2,1)
AC067959.1(6,0,4,0)
BCL6(11,5,11,2)
LPP(2,0,3,1)
DLG1(4,1,0,1)
PIM1(2,1,3,0)
HCG23(2,0,3,0)
ATG5(5,0,0,1)
TOX(2,1,1,3)
CD274(0,4,0,1)
MTAP(20,2,4,5)
CDKN2A(5,0,0,1)
CDKN2B−AS1(9,1,2,0)
RP11−408N14.1(3,0,0,2)
RP11−266K4.9(3,0,1,1)
OVOS(0,0,3,0)
RP11−726G1.2(0,0,3,0)
ETV6(9,2,6,1)
AL021546.6(5,0,0,0)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

Figure S35

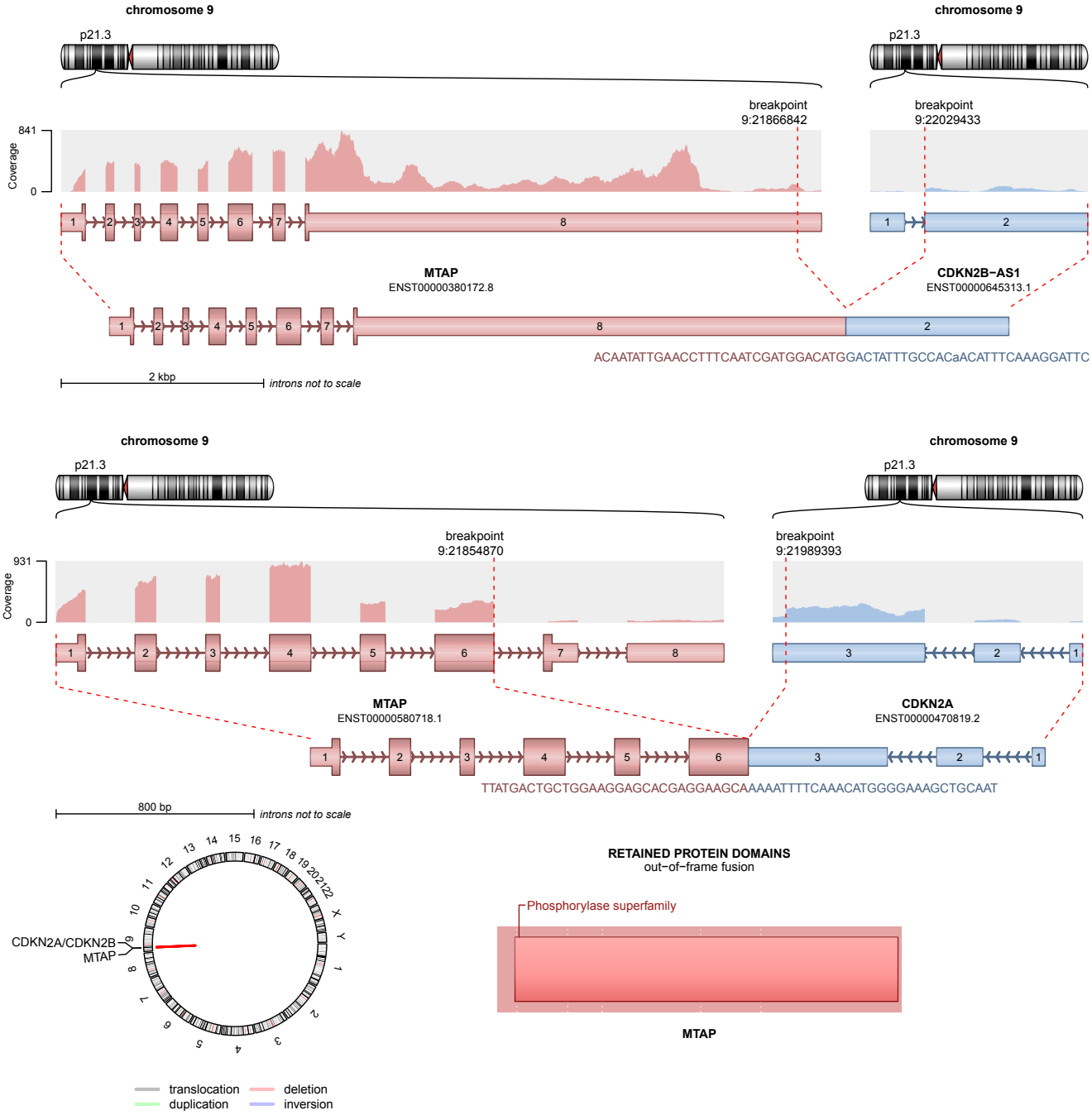

chromosome 9
p21.3
breakpoint
9:21866842
breakpoint
9:22029433
841
Coverage
0
MTAP
ENST00000380172.8
CDKN2B−AS1
ENST00000645313.1
ACAATATTGAACCTTTCAATCGATGGACATGGACTATTTGCCACaACATTTCAAAGGATTC
2 kbp
introns not to scale
chromosome 9
p21.3
breakpoint
9:21854870
breakpoint
9:21989393
931
Coverage
0
MTAP
ENST00000580718.1
CDKN2A
ENST00000470819.2
TTATGACTGCTGGAAGGAGCACGAGGAAGCAAAAATTTTCAAACATGGGGAAAGCTGCAAT
800 bp
introns not to scale
CDKN2A/CDKN2B
MTAP
translocation
deletion
duplication
inversion
RETAINED PROTEIN DOMAINS
out−of−frame fusion
Phosphorylase superfamily
MTAP

Figure S36

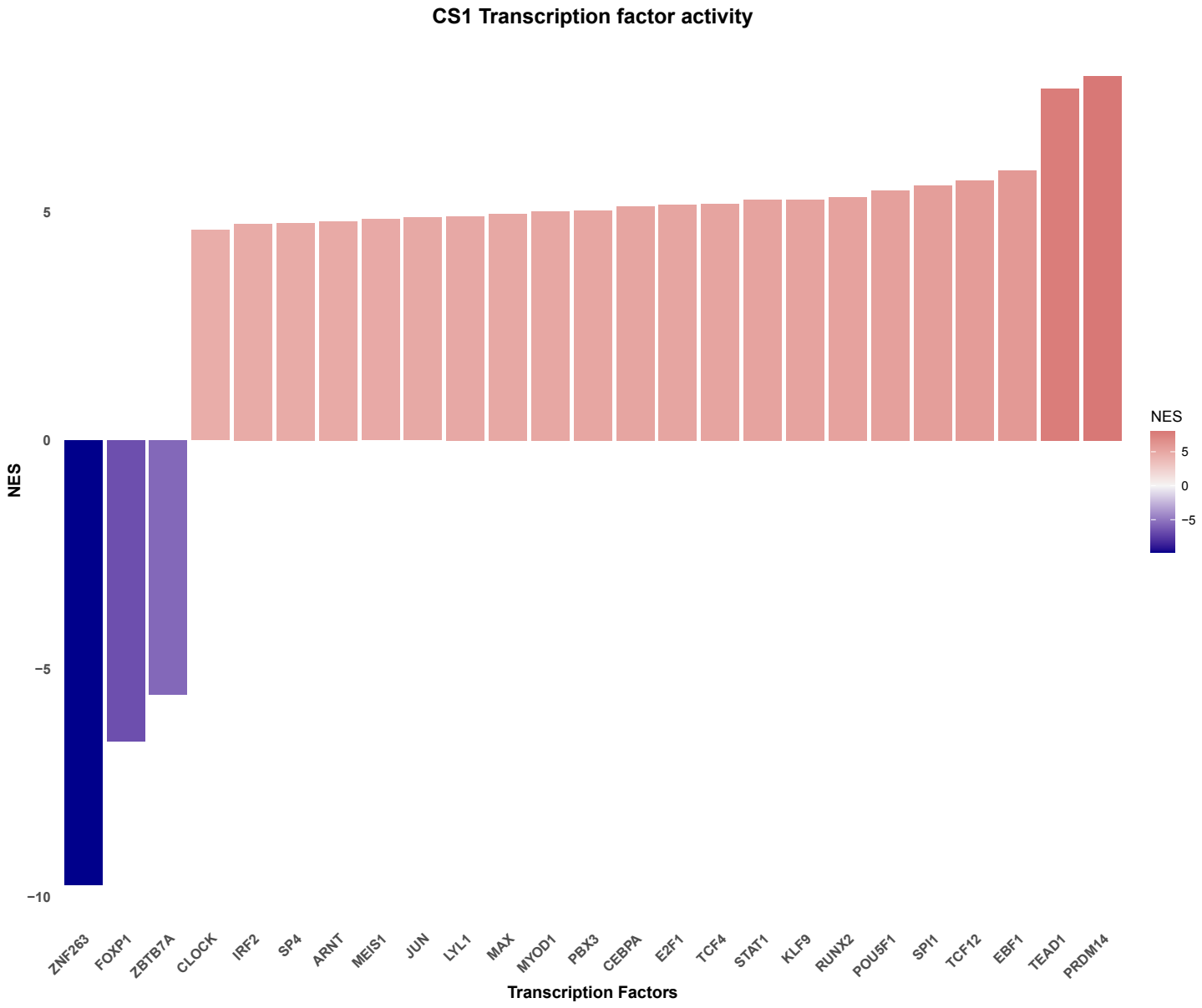

CS1 Transcription factor activity
NES
5
0
−5
−10
ZNF263
FOXP1
ZBTB7A
CLOCK
IRF2
SP4
ARNT
MEIS1
JUN
LYL1
MAX
MYOD1
PBX3
CEBPA
E2F1
TCF4
STAT1
KLF9
RUNX2
POU5F1
SPI1
TCF12
EBF1
TEAD1
PRDM14
Transcription Factors
NES
5
0
−5

Figure S37

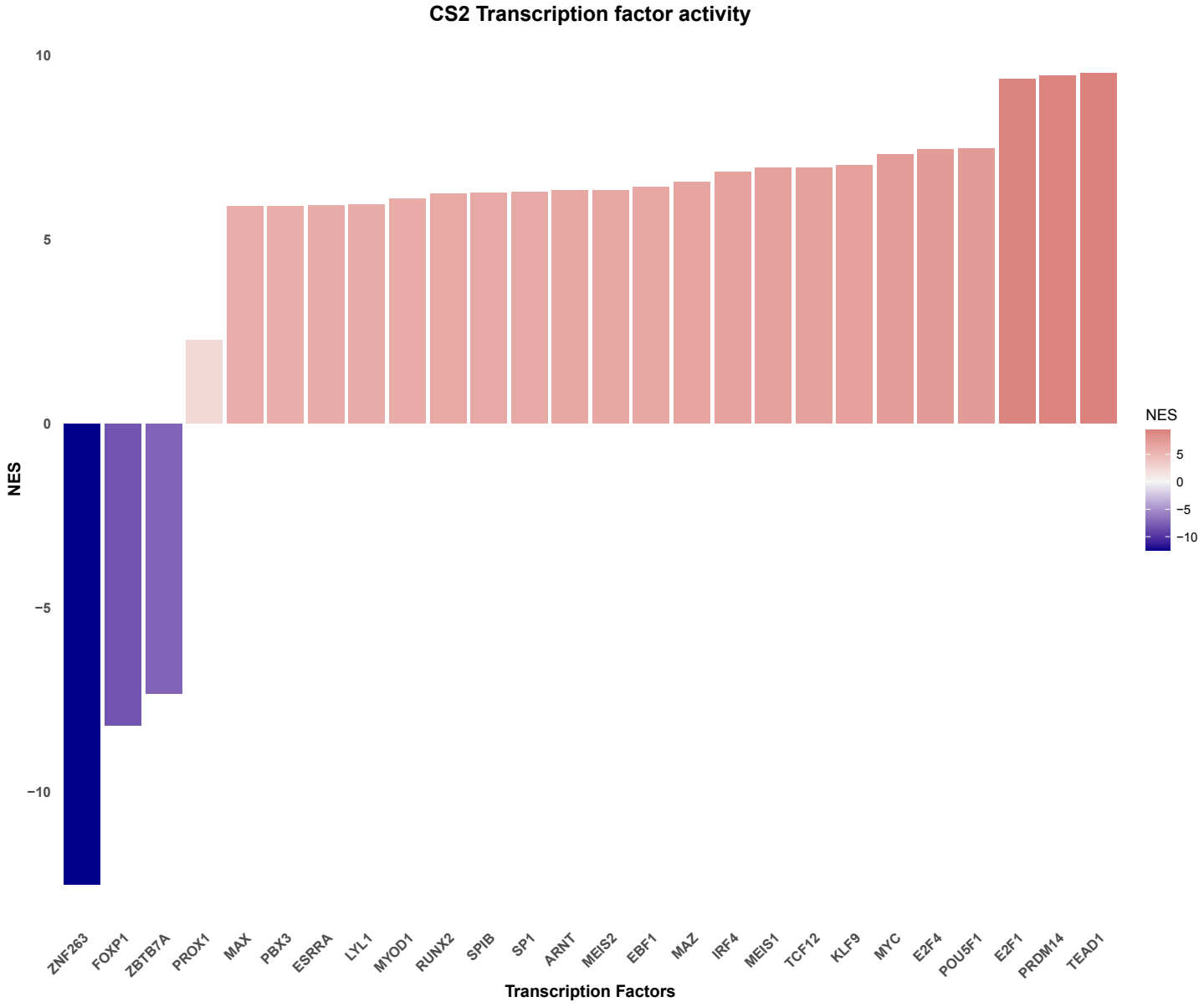

CS2 Transcription factor activity
10
5
0
−5
−10
NES
NES
5
0
−5
−10
ZNF263
FOXP1
ZBTB7A
PROX1
MAX
PBX3
ESRRA
LYL1
MYOD1
RUNX2
SPIB
SP1
ARNT
MEIS2
EBF1
MAZ
IRF4
MEIS1
TCF12
KLF9
MYC
E2F4
POU5F1
E2F1
PRDM14
TEAD1
Transcription Factors

Figure S38

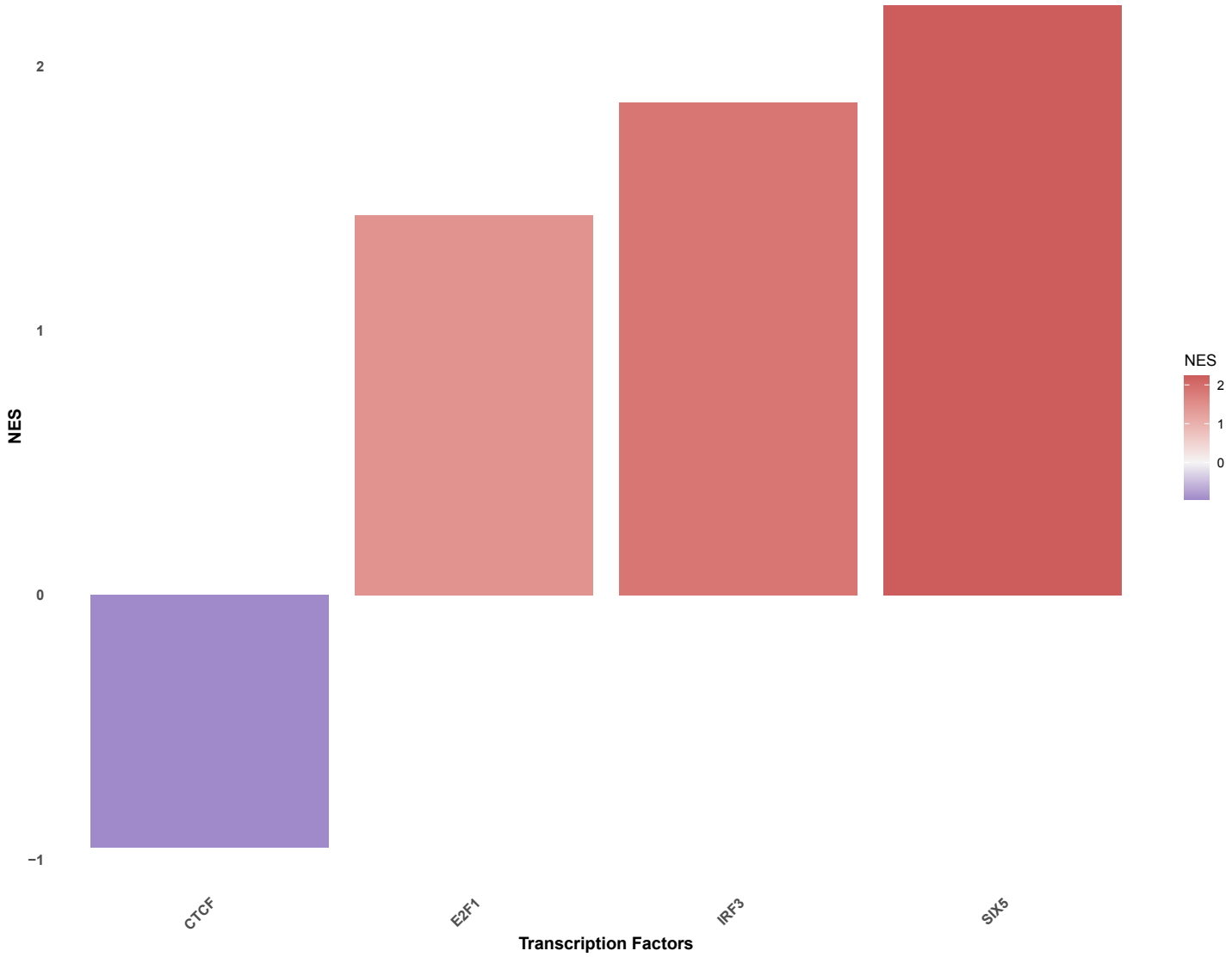

CS3 Transcription factor activity
NES
2
1
0
−1
CTCF
E2F1
IRF3
SIX5
Transcription Factors
NES
2
1
0

Figure S39

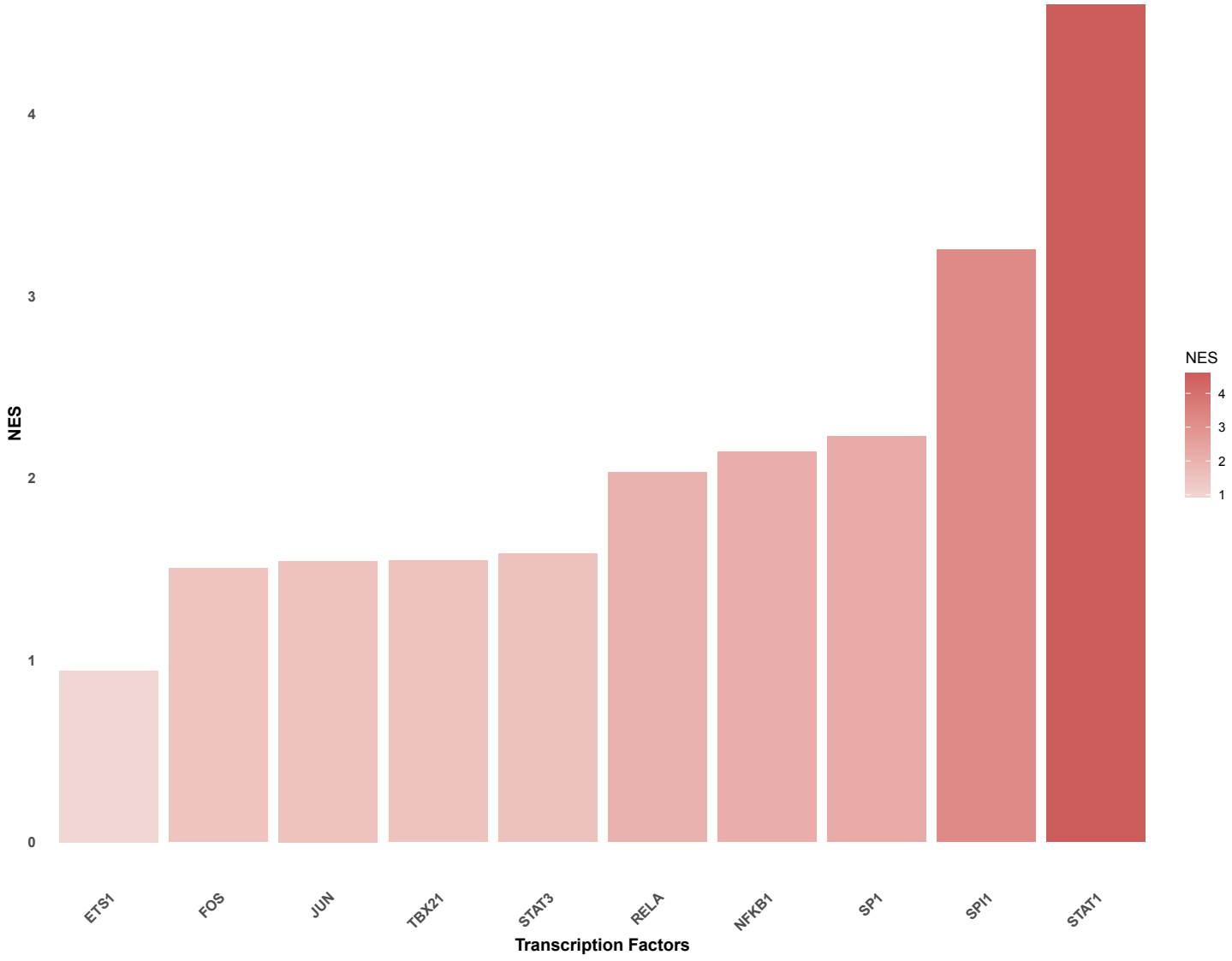

CS4 Transcription factor activity
NES
4
3
2
1
0
ETS1
FOS
JUN
TBX21
STAT3
RELA
NFKB1
SP1
SPI1
STAT1
Transcription Factors
NES
4
3
2
1

Figure S40

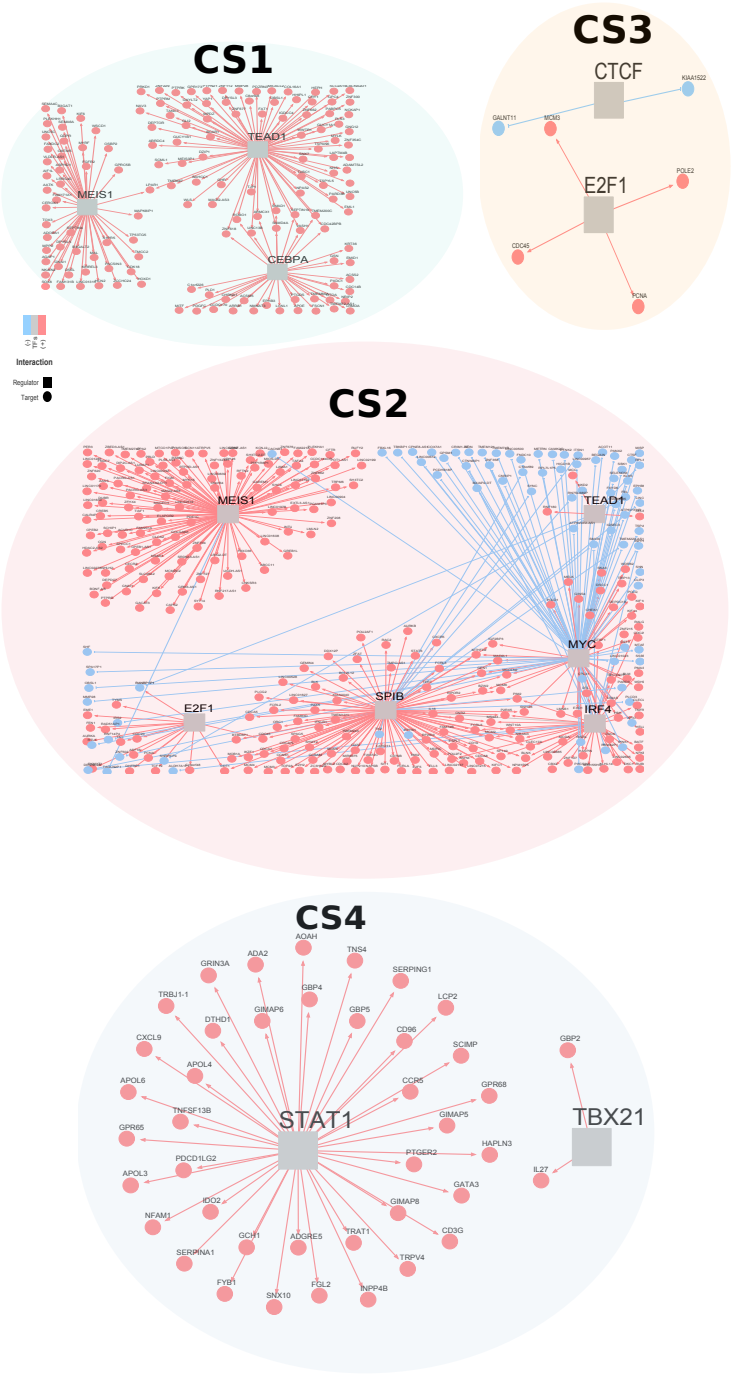

CS1
TEAD1
MEIS1
CEBPA
CS3
CTCF
KIAA1522
GALNT11
MCM3
E2F1
POLE2
CD45
PCNA
Interaction
Regulator
Target
CS2
MEIS1
TEAD1
MYC
SPIB
E2F1
IRF4
CS4
AOAH
TNS4
ADA2
GRIN3A
SERPING1
GBP4
TRBJ1-1
LCP2
GIMAP6
GBP5
DTHD1
CD96
CXCL9
SCIMP
GBP2
APOL4
APOL6
CCR5
GPR68
TNFSF13B
STAT1
GIMAP5
TBX21
GPR65
HAPLN3
PTGER2
PDCD1LG2
IL27
APOL3
GATA3
IDO2
GIMAP8
NFAM1
GCH1
ADGRE5
TRAT1
CD3G
SERPINA1
TRPV4
FYB1
FGL2
INPP4B
SNX10

Figure S41

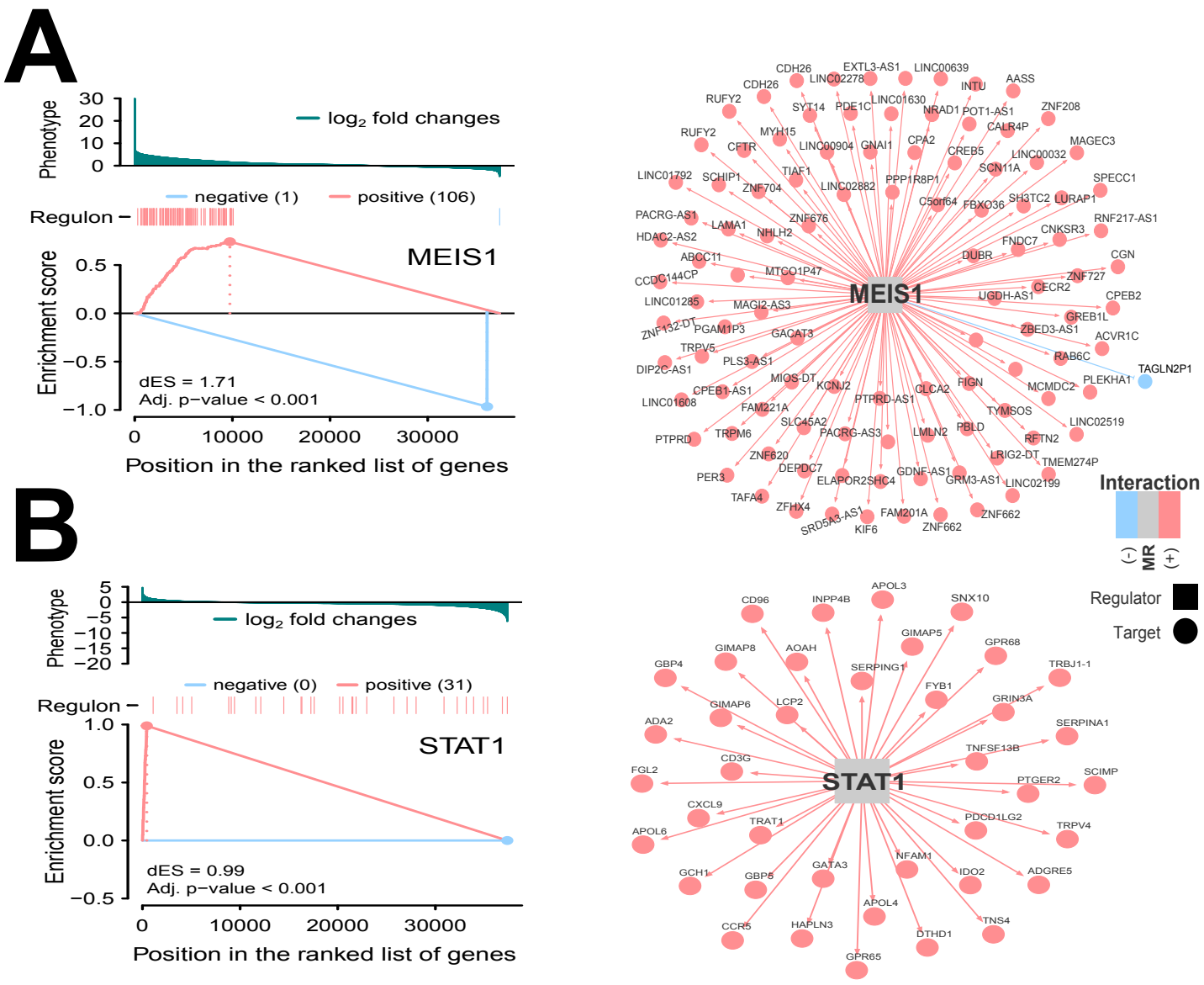

A
Phenotype
log2 fold changes
negative (1)
positive (106)
Regulon
MEIS1
Enrichment score
dES = 1.71
Adj. p−value < 0.001
Position in the ranked list of genes
B
Phenotype
log2 fold changes
negative (0)
positive (31)
Regulon
STAT1
Enrichment score
dES = 0.99
Adj. p−value < 0.001
Position in the ranked list of genes
MEIS1
STAT1
Interaction
MR
Regulator
Target

Figure S42

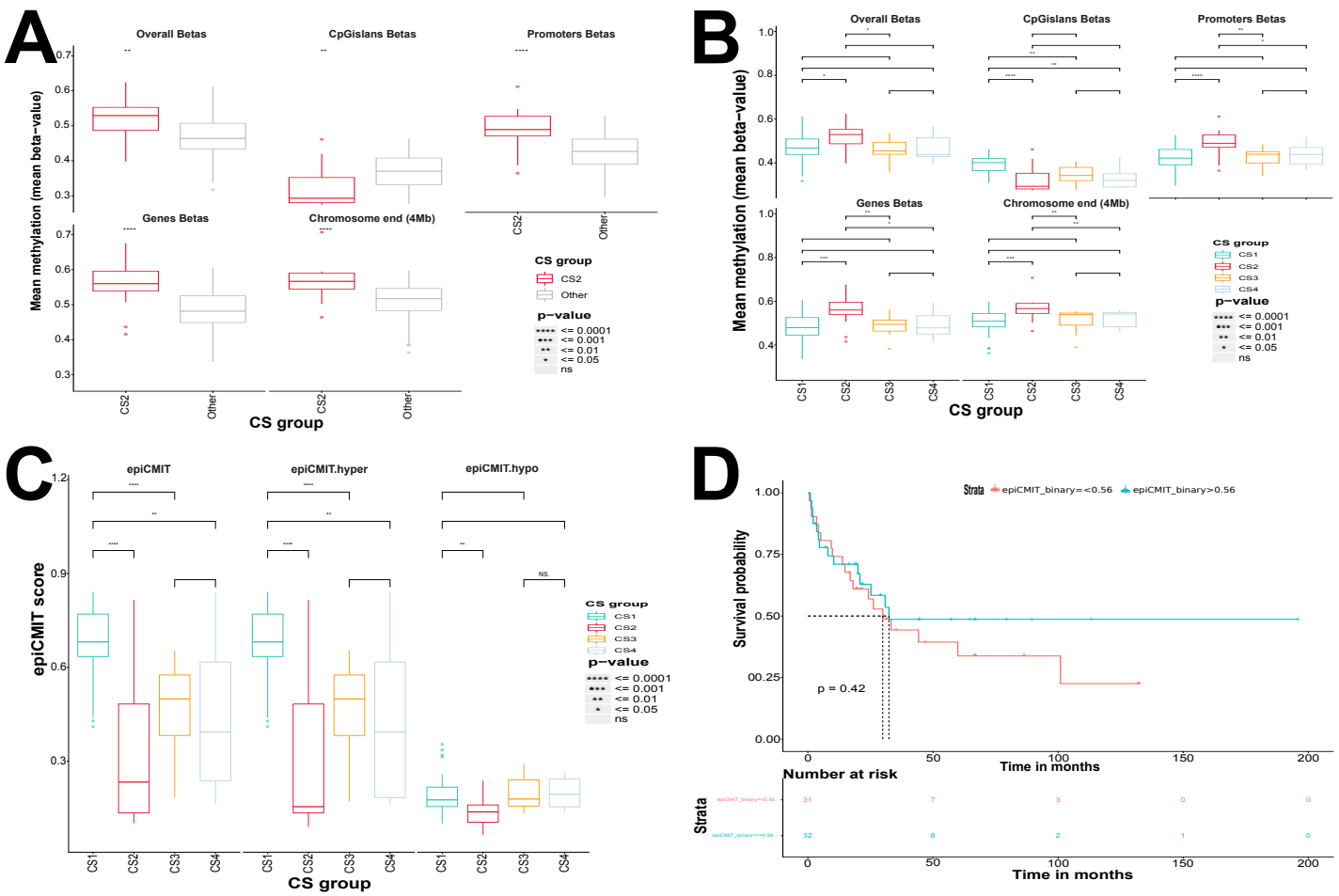

A
Overall Betas
CpGislans Betas
Promoters Betas
Genes Betas
Chromosome end (4Mb)
Mean methylation (mean beta−value)
CS group
CS2
Other
p−value
**** <= 0.0001
*** <= 0.001
** <= 0.01
* <= 0.05
ns
B
Overall Betas
CpGislans Betas
Promoters Betas
Genes Betas
Chromosome end (4Mb)
Mean methylation (mean beta−value)
CS group
CS1
CS2
CS3
CS4
p−value
**** <= 0.0001
*** <= 0.001
** <= 0.01
* <= 0.05
ns
C
epiCMIT
epiCMIT.hyper
epiCMIT.hypo
epiCMIT score
NS.
CS group
CS1
CS2
CS3
CS4
p−value
**** <= 0.0001
*** <= 0.001
** <= 0.01
* <= 0.05
ns
D
Strata
epiCMIT_binary=<0.56
epiCMIT_binary>0.56
Survival probability
p = 0.42
Time in months
Number at risk
Strata
31
7
3
0
0
32
8
2
1
0
Time in months

Figure S43

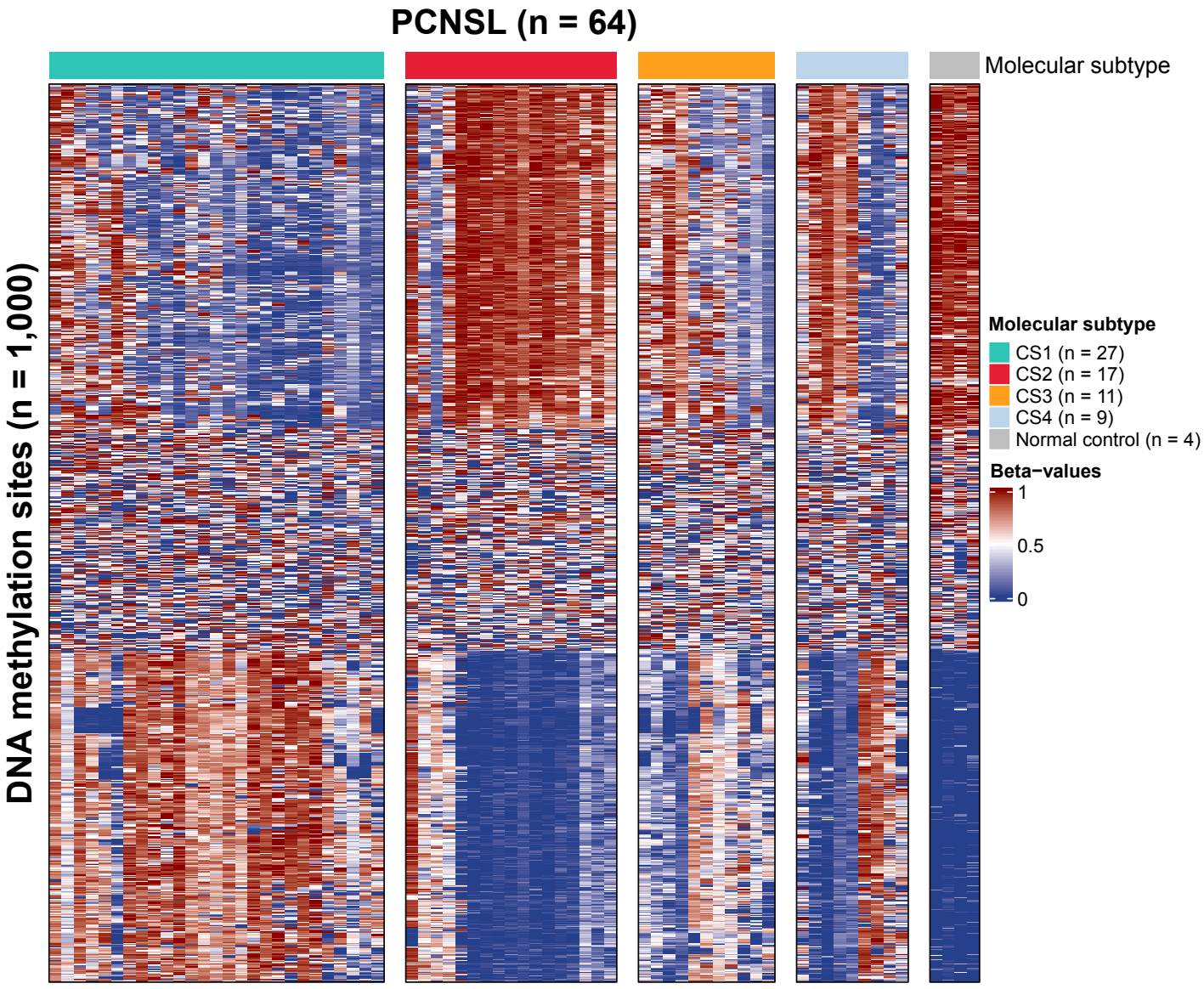
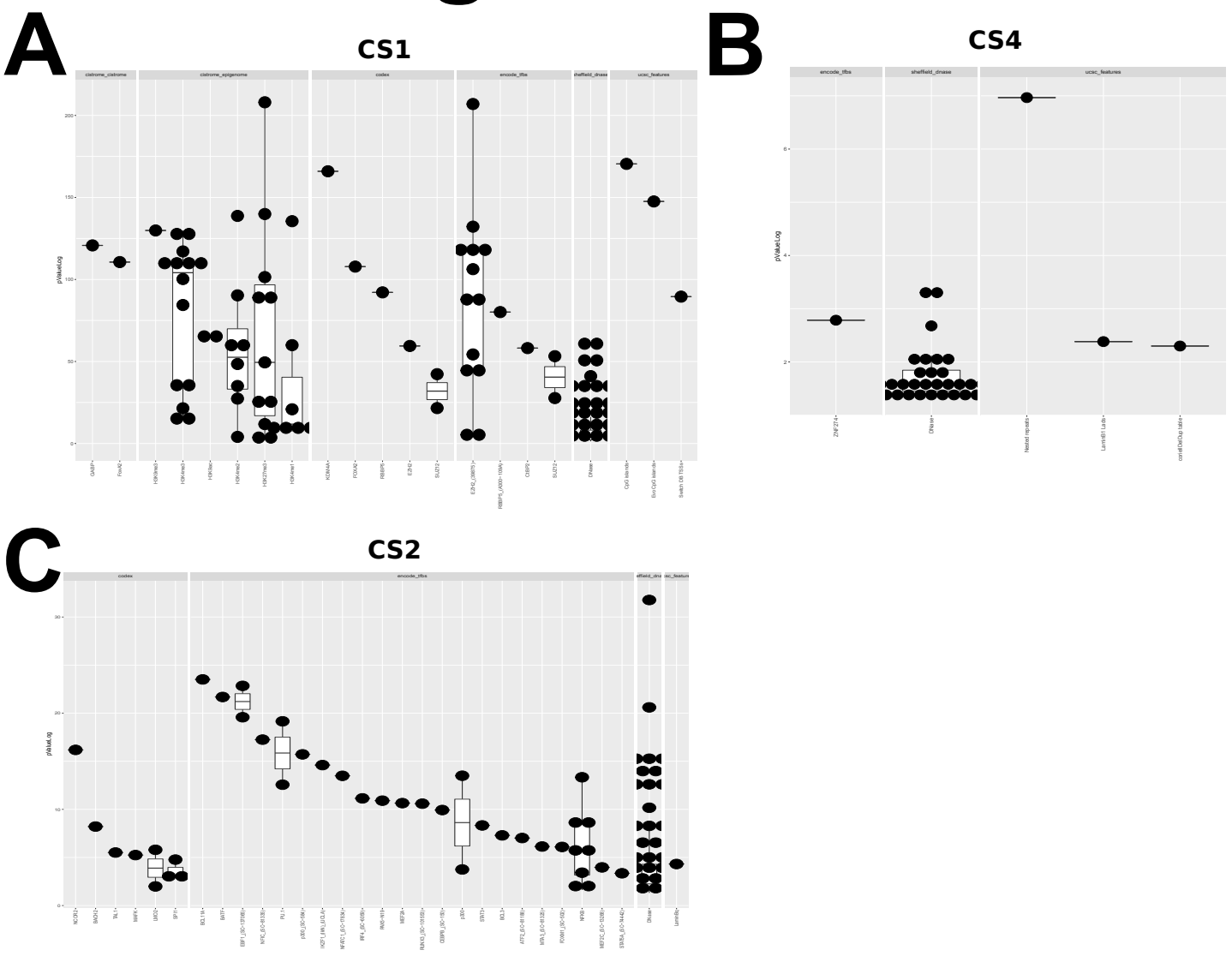

PCNSL (n = 64)
DNA methylation sites (n = 1,000)
Molecular subtype
Molecular subtype
CS1 (n = 27)
CS2 (n = 17)
CS3 (n = 11)
CS4 (n = 9)
Normal control (n = 4)
Beta−values
1
0.5
0

Figure S44



Second Appendix

Additional publications:

1. **HERNANDEZ Isaias**, ALENTORN Agusti (**2021**). ACTIVATION-INDUCED CYTIDINE DEAMINASE AS A NEW BIOMARKER (European Patent Application No. EP21305876.1). **European Patent Office**.

2. Ytel Garcilazo-Reyes, Maria-José Ibáñez-Juliá, **Isaias Hernández-Verdin**, Ludovic Nguyen-Them, Nadia Younan, Caroline Houillier, Khê Hoang-Xuan, and Agusti Alentorn (**2020**).Treating central nervous system lymphoma in the era of precision medicine. **Expert Review of Precision Medicine and Drug Development**.

3. Mohammed H. Ahmed, **Isaias Hernández-Verdin**, Nolwenn Lemaire, Emie Quissac, Coralie L Guerin, Lea Guyonnet, Noël Zahr, Laura Mouton, Mathieu Santin, Alexandra Petiet, Charlotte, Schmitt, Guillaume Bouchoux, Michael Canney, Marc Sanson, Maite Verreault, Alexandre Carpentier, Ahmed Idbaih (**2021**). Increased brain delivery of anti-programmed death-ligand 1 using low-intensity pulsed ultrasound-mediated blood-brain barrier opening is associated with increased anti-tumor efficacy and microglia activation in glioblastoma mouse models. **Neuro-Oncology** (Under revision).

4. Mohammed H. Ahmed, **Isaias Hernández-Verdin**, Franck Bielle, Maite Verreault, Julie Lerond, Agusti Alentorn, Marc Sanson, and Ahmed Idbaih (**2022**). Expression and Prognostic Value of CD80 and CD86 in the Tumor Microenvironment of Newly Diagnosed Glioblastoma. **Canadian Journal of Neurological Sciences**.

Europäisches Patentamt | European Patent Office | Office européen des brevets

Acknowledgement of receipt

We hereby acknowledge receipt of your request for grant of a European patent as follows:

Submission number	1000502871	
Application number	EP21305876.1	
File No. to be used for priority declarations	EP21305876	
Date of receipt	24 June 2021	
Your reference	IBIO-1905/EP	
Applicant	ICM (INSTITUT DU CERVEAU ET DE LA MOELLE ÉPINIÈRE)	
Country	FR	
Title	ACTIVATION-INDUCED CYTIDINE DEAMINASE AS A NEW BIOMARKER	
Documents submitted	package-data.xml	ep-request.xml
	application-body.xml	ep-request.pdf (5 p.)
	SPECEPO-1.pdf\1905 EP.pdf (57 p.)	f1002-1.pdf (1 p.)
Submitted by	EMAIL=naj@icosa.fr,CN=Nathalie JOUANNIC,O=CABINET ICOSA,C=FR	
Method of submission	Online	
Date and time receipt generated	24 June 2021, 21:54:05 (CEST)	
Official Digest of Submission	8D:78:15:FE:72:90:CF:50:98:7D:BF:5A:65:9D:E4:8C:83:28:F4:08	

/INPI, section dépôt/

Request for grant of a European patent

For official use only		
1 Application number:	MKEY	
2 Date of receipt (Rule 35(2) EPC):	DREC	
3 Date of receipt at EPO (Rule 35(4) EPC):	RENA	
4 Date of filing:		

5 Grant of European patent, and examination of the application under Article 94, are hereby requested. ☒

5.1 The applicant waives his right to be asked whether he wishes to proceed further with the application (Rule 70(2)) ☐

Procedural language: en

Filing Language: en

6 Applicant's or representative's reference: IBIO-1905/EP

Filing Office: FR

Applicant 1

7-1 Name: ICM (INSTITUT DU CERVEAU ET DE LA MOELLE ÉPINIÈRE)

8-1 Address: Hôpital Pitié Salpêtrière
47/83 boulevard de l'Hôpital
75013 Paris
France

10-1 State of residence or of principal place of business: France

Applicant 2

7-2 Name: INSERM (Institut National de la Santé et de la Recherche Médicale)

8-2 Address: 101 rue de Tolbiac
75654 Paris Cedex 13
France

10-1 State of residence or of principal place of business: France

Applicant 3

7-3	Name:	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
8-3	Address:	3 rue Michel-Ange 75016 Paris France
10-1	State of residence or of principal place of business:	France

Applicant 4

7-4	Name:	APHP (Assistance Publique - Hôpitaux de Paris)
8-4	Address:	3 avenue Victoria 75004 Paris France
10-1	State of residence or of principal place of business:	France

Applicant 5

7-5	Name:	SORBONNE UNIVERSITE
8-5	Address:	21 rue de l'École de Médecine 75006 Paris France
10-1	State of residence or of principal place of business:	France

14.1 The/Each applicant hereby declares that he is an entity or a natural person under Rule 6(4) EPC. ☐

Representative 1

15-1	Name:	ICOSA
	Association No.:	379
16-1	Address of place of business:	83 avenue Denfert-Rochereau 75014 Paris France
17-1	Telephone:	+33 (0)1 53 63 46 90
17-1	Fax:	+33 (0)1 42 84 46 50
17-1	E-mail:	icosapatent@icosa.fr

Inventor(s)

23 Designation of inventor attached ☒

24 Title of invention

Title of invention: ACTIVATION-INDUCED CYTIDINE DEAMINASE AS A NEW BIOMARKER

25 Declaration of priority (Rule 52) and search results under Rule 141(1)

A declaration of priority is hereby made for the following applications

25.2 Re-establishment of rights

Re-establishment of rights under Article 122 EPC in respect of the priority period is herewith requested for the following priority/priorities

25.3 The EPO is requested to retrieve a certified copy of the following previous application(s) (priority document(s)) via the WIPO Digital Access Service (DAS) using the indicated access code(s):

Request	Application number:	Access Code

25.4 This application is a complete translation of the previous application ☐

25.5 It is not intended to file a (further) declaration of priority ☐

26 Reference to a previously filed application

27 Divisional application ☐

28 Article 61(1)(b) application ☐

29 Claims

Number of claims: 13

29.1 ☒ as attached

29.2 ☐ as in the previously filed application (see Section 26.2)

29.3 ☐ The claims will be filed later

30 Figures

It is proposed that the abstract be published together with figure No. 1

31 Designation of contracting states

All the contracting states party to the EPC at the time of filing of the European patent application are deemed to be designated (see Article 79(1)).

32 Different applicants for different contracting states

33 Extension/Validation

This application is deemed to be a request to extend the effects of the European patent application and the European patent granted in respect of it to all non-contracting states to the EPC with which extension or validation agreements are in force on the date on which the application is filed. However, the request is deemed withdrawn if the extension fee or the validation fee, whichever is applicable, is not paid within the prescribed time limit.

33.1 It is intended to pay the extension fee(s) for the following state(s):

33.2 It is intended to pay the validation fee(s) for the following state(s):

34 Biological material

38 Nucleotide and amino acid sequences

38.1 The description contains a sequence listing. ☐

38.2a The sequence listing is attached in computer-readable format in accordance with WIPO Standard ST.25 (Rule 30(1)). ☐

38.2b The sequence listing is attached in PDF format ☐

Further indications

39 Additional copies of the documents cited in the European search report are requested

Number of additional sets of copies:

40 Refund of the search fee under Article 9(2) of the Rules relating to Fees is requested ☐

Application number or publication number of earlier search report:

42 Payment

Method of payment: Automatic debit order

The European Patent Office is hereby authorised, under the Arrangements for the automatic debiting procedure, to debit from the deposit account any fees and costs falling due.

Currency: EUR

Deposit account number: 28040791

Account holder: ICOSA

43 Refunds

Any refunds should be made to EPO deposit account: 28040791

Account holder: ICOSA

44-A	**Forms**	Details:	System file name:
A-1	Request		as ep-request.pdf
A-2	1. Designation of inventor	1. Inventor	as f1002-1.pdf

44-B	**Technical documents**	Original file name:	System file name:
B-1	Specification	1905 EP.pdf Description; 13 claims; 17 figure(s); abstract	SPECEPO-1.pdf

44-C	**Other documents**	Original file name:	System file name:

Annotations

Title (Author):

1Note (for EPO)

A l'attention d'INPI (ICOSA)

Nous autorisons l'INPI à prélever sur notre compte INPI n°3924 la taxe de transmission à l'OEB et toutes taxes nécessaires à la bonne exécution du dépôt.

45

General authorisation:

46 Signature(s)

Place:	**Lyon**
Date:	**24 June 2021**
Signed by:	**FR, CABINET ICOSA, Nathalie JOUANNIC**
Association:	**ICOSA**
Representative name:	**JOUANNIC**
Capacity:	**(Representative)**

Form 1002 - 1: Public inventor(s)

Designation of inventor

User reference: IBIO-1905/EP
Application No:

Public

	Inventor	
	Name:	ALERTORN Augusti
	Address:	75013 Paris France
	The applicant has acquired the right to the European patent:	As employer
	Inventor	
	Name:	HERNANDEZ Isaias
	Address:	75014 PARIS France
	The applicant has acquired the right to the European patent:	As employer

Signature(s)

Place:	**Lyon**
Date:	**24 June 2021**
Signed by:	**FR, CABINET ICOSA, Nathalie JOUANNIC**
Association:	**ICOSA**
Representative name:	**JOUANNIC**
Capacity:	**(Representative)**

1

ACTIVATION-INDUCED CYTIDINE DEAMINASE AS A NEW BIOMARKER

FIELD OF INVENTION

[1] The present invention relates to an activation-induced cytidine deaminase (AICDA) as a new biomarker for cancer.

BACKGROUND OF INVENTION

[2] According to the World Health Organization, a biomarker is any substance, structure or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease. In the field of cancer, a biomarker can be used for assessing multiple factors including determining the risk of developing cancer, monitoring cancer progression, determining the survival prognosis of cancer patients, or predicting potential response to therapy. Therefore, robust cancer biomarkers are increasingly needed.

[3] Activation-induced cytidine deaminase (AID, encoded by *AICDA*) belongs to a large family of enzymes called apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) which are considered as a source of somatic mutations in the genome. AID was initially described as the driver of somatic hypermutation (SHM), which diversifies the variable (V) domains of immunoglobulin genes in activated B cells in germinal centers. Furthermore, AID is also involved in class-switch recombination (CSR) in immunoglobulins. These processes are related to the conversion of deoxycytidines into deoxyuridines.

[4] The present invention describes the use of AID-related mutations as a new biomarker for cancer. More specifically, the present invention relates to the use of AID-related mutations for prognosing survival in cancer patients treated with immune checkpoint inhibitors, but also to identify cancer patients susceptible to respond to immune checkpoint inhibitor therapy.

SUMMARY

[5] The present invention relates to an *in vitro* method for identifying a subject with cancer as being susceptible to respond to a treatment with an immune checkpoint inhibitor (ICI) or for prognosing survival of a subject with cancer and being treated with ICI, the method comprising assessing the fraction of AID-related mutations in a sample, wherein the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants.

[6] In one embodiment, said AID-related mutation is a mutation falling into an AID hotspot sequence, wherein said AID hotspot sequence includes the nucleic sequence WRCY or its reverse RGYW.

[7] In one embodiment, said AID hotspot sequence includes AACC, AACT, AGCC, AGCT, TACC, TACT, TGCC, TGCT or the reverse TTGG, TTGA, TCGG, TCGA, ATGG, ATGA, ACGG, ACGA.

[8] In one embodiment, said sample is a tumor tissue.

[9] In one embodiment, said method is for identifying a subject with cancer as being susceptible to respond to a treatment with an ICI.

[10] In one embodiment, said method is for prognosing survival of a subject with cancer and being treated with ICI.

[11] In one embodiment, the ICI is selected from the group comprising an inhibitor of PD-1, an inhibitor of PD-L1, an inhibitor of CTLA-4 and a combination thereof.

[12] In one embodiment, the cancer is selected from the group comprising melanoma, non-small-cell lung carcinoma (NSCLC), renal cell carcinoma, head and neck cancers, merkel-cell carcinoma, gastric cancer, small-cell lung carcinoma (SCLC), Hodgkin lymphoma, breast cancer, cervical cancer, colorectal cancer, endometrial cancer, hepatocellular cancer, esophageal cancer, mesothelioma, MSI (microsatellite instability)-high solid tumors, TMB (tumor mutation burden)-high tumors, breast cancer and urothelial carcinoma.

[13] In one embodiment, said method further comprising the step of comparing the fraction of AID-related mutations with a reference value.

[14] In one embodiment, said reference value is the median of the fractions of AID-related mutations measured in a reference population.

[15] In one embodiment, said reference value is a decile of the fractions of AID-related mutations measured in a reference population.

[16] In one embodiment, the reference population is a population of subjects having or having had a cancer, which are or have been treated with ICI, and which respond or have responded to ICI.

[17] In one embodiment, a fraction of AID-related mutation above the reference value is indicative of a subject as being susceptible to respond to a treatment with ICI, or prognosed with a high survival.

DEFINITIONS

[18] In the present invention, the following terms have the following meanings:

[19] "**About**" preceding a figure encompasses plus or minus 10%, or less, of the value of said figure. It is to be understood that the value to which the term "about" refers is itself also specifically, and preferably, disclosed.

[20] "**Activation-induced cytidine deaminase**" or "**AID**" belongs to a large family of enzymes called apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC), encoded by *AICDA*.

[21] "**Comprise**" is intended to mean "**contain**", "**encompass**" and "**include**". In some embodiments, the term "**comprise**" also encompasses the term "**consist of**".

[22] "**Biomarker**" refers to any substance, structure or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease.

[23] "**Quantile(s)**" refers to (a) cut-off value(s) dividing the observations/measures made in a population into equal-sized groups, each group comprising an equal percentage of said observations/measures. As used herein, "quantile(s)" thus refer to cut-off fraction

of AID-related mutations dividing the fractions of AID-related mutations measured in a biological sample from each of the human subjects of a reference population into equal-sized groups each comprising an equal percentage of said measures of fractions of AID-related mutations. In other words, "quantile(s)" refers to cut-off fractions of AID-related mutations below or above which lies a determined percentage of the fractions of AID-related mutations measured in a reference population. For example, as used herein, "median" refer to the cut-off fraction of AID-related mutations dividing the fractions of AID-related mutations measured in a reference population into two groups, each comprising 50% of the fractions of AID-related mutations measured in the reference population. It should be noted that "quantiles" may also sometimes refer to the groups so defined by said cut-off value. For example, "median" may also sometimes refer to the two groups defined by the cut-off fraction of AID-related mutations dividing the fractions of AID-related mutations measured in a reference population. However, as used herein and unless otherwise specified, the term "quantile" refers to a cut-off value.

[24] "**Prognosis**": refers to the likely outcome or course of a disease, for example a cancer; the chance of remission or recurrence.

[25] "**Single nucleotide variant**": refers to a variation in a single nucleotide.

[26] "**Subject**": refers to an animal, preferably a mammalian subject, more preferably a human subject. Among the non-human mammalian subjects of interest, one may non-limitatively mention pets, such as dogs, cats, guinea pigs; animals of economic importance such as cattle, sheep, goats, horses, monkeys. In one embodiment, a subject may be a "patient", *i.e.* a warm-blooded animal, more preferably a human, who/which is awaiting the receipt of, or is receiving medical care or was/is/will be the object of a medical procedure, or is monitored for the development of a disease, disorder or condition. In one embodiment, the subject is an adult (for example a human subject above the age of 18). In another embodiment, the subject is a child (for example a human subject below the age of 18). In one embodiment, the subject is a male. In another embodiment, the subject is a female.

[27] **"Susceptible to respond to a treatment"** refers to the likelihood of a response to the treatment.

[28] **"Treating"** or **"treatment"** or **"alleviation"** refers to both therapeutic treatment and prophylactic or preventative measures, wherein the object is to prevent or slow down (lessen) cancer; Those in need of treatment include those already with cancer, as well as those prone to develop cancer or those in whom cancer is to be prevented. An individual is successfully "treated" for cancer, if, after receiving a therapeutic amount of the immune checkpoint inhibitor, the individual shows observable and/or measurable reduction in or absence of one or more of the symptoms associated with cancer; reduced morbidity and mortality, and improvement in quality of life issues. The above parameters for assessing successful treatment and improvement of cancer are readily measurable by routine procedures familiar to physician or authorized personnel.

[29] **"Therapeutically efficient amount**" refers to the level or the amount of the active agent, in particular an immune checkpoint inhibitor, which is aimed at, without causing significant negative or adverse side effects to the target, (1) delaying or preventing the onset of cancer; (2) slowing down or stopping the progression, aggravation, or deterioration of one or more symptoms of cancer; (3) bringing about amelioration of the symptoms of cancer; (4) reducing the severity or incidence of cancer; or (5) curing cancer. A therapeutically effective amount may be administered prior to the onset of cancer, for a prophylactic or preventive action. Alternatively, or additionally, the therapeutically effective amount may be administered after the onset of cancer for a therapeutic action. In one embodiment, a therapeutically effective amount of the pharmaceutical composition is an amount that is effective in reducing at least one symptom of cancer.

DETAILED DESCRIPTION

[30] The present invention relates to an *in vitro* method for identifying a subject with cancer as being susceptible to respond to a treatment with an immune checkpoint inhibitor (ICI) or for prognosing survival of a subject with cancer and being treated with ICI, the method comprising assessing the fraction of AID-related mutations in a sample, wherein the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants.

[31] In one embodiment, the AID-related mutation is a mutation falling into an AID hotspot sequence.

[32] In one embodiment, the number of AID-related mutations is the number of mutated cytosines (and guanines) falling into an AID hotspot sequence. In one embodiment, the mutated cytosines (and guanines) are converted into uraciles.

[33] In one embodiment, the AID hotspot sequence includes the nucleic sequence WRCY or its reverse RGYW.

[34] According to the invention, R is purine, G is Guanine (G), C is Cytosine (C), Y is pyrimidine, and W is Adenine (A) or Thymine (T). According to the invention, purines include Adenine and Guanine, while pyrimidines include Cytosine, Uracile, and Thymine.

[35] In one embodiment, the AID hotspot sequence includes AACC, AACT, AGCC, AGCT, TACC, TACT, TGCC, TGCT for the positive strand and TTGG, TTGA, TCGG, TCGA, ATGG, ATGA, ACGG, ACGA for the negative strand.

[36] In one embodiment, the AID-related mutation is located in tumours, *i.e.* in tumoral cells.

[37] In one embodiment, the AID-related mutation can be located in any genomic region. In one embodiment, the AID-related mutations are in a higher density in chromosome 5, chromosome 17 and/or chromosome 2 of the subject.

7

[38] In one embodiment, the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants measured in a sample.

[39] In one embodiment, the sample is a tumor tissue. In one embodiment, the tumor tissue is obtained by a biopsy.

[40] In one embodiment, the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants measured in the whole genome of the tumoral cell. In one embodiment, the whole genome sequence is obtained by whole genome sequencing (WGS).

[41] In one embodiment, the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants measured in the whole exome of the tumoral cell. In one embodiment, the whole exome sequence is obtained by whole exome sequencing (WES).

[42] In one embodiment, the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants measured in a set of around 400 genes associated with tumors of the tumoral cell. In one embodiment, the sequence of the set of genes is obtained by targeted next generation sequencing (NGS).

[43] In one embodiment, the method as described hereinabove is for identifying a subject with cancer as being susceptible to respond to a treatment with an immune checkpoint inhibitor.

[44] In one embodiment, the cancer is a solid cancer. As used herein, a solid cancer encompasses cancers that form an abnormal mass of tissues that usually does not contain cysts or liquid areas. Different types of solid tumors are named for the type of cells that form them. Examples of solid tumors are sarcomas, carcinomas, and lymphomas.

[45] In one embodiment, the cancer is a blood cancer. As used herein, a blood cancer or a hematological cancer is a cancer that begins in blood-forming tissue, such as the bone

marrow, or in the cells of the immune system. Examples of blood cancer are leukemia, lymphoma, and multiple myeloma.

[46] In one embodiment, the cancer is selected form the group comprising melanoma, non-small-cell lung carcinoma (NSCLC), renal cell carcinoma, head and neck cancers, merkel-cell carcinoma, gastric cancer, small-cell lung carcinoma (SCLC), Hodgkin lymphoma, breast cancer, cervical cancer, colorectal cancer, endometrial cancer, hepatocellular cancer, esophageal cancer, mesothelioma, MSI (microsatellite instability)-high solid tumors, TMB (tumor mutation burden)-high tumors, breast cancer and urothelial carcinoma.

[47] In one embodiment, the cancer is melanoma, preferentially metastatic melanoma.

[48] In one embodiment, the cancer is a cancer with unknown primary. As used herein, cancers of unknown primary (CUPs) are histologically confirmed, metastatic malignancies with a primary tumor site that is unidentifiable on the basis of standard evaluation and imaging studies. Standard evaluation and imaging studies for detecting tumors are well-known by the skilled artisan in the art.

[49] In one embodiment, the cancer is a carcinoma, preferentially a renal cell carcinoma.

[50] In one embodiment, the subject suffers from a cancer.

[51] In one embodiment, the subject has been previously diagnosed with cancer by authorized personnel skilled in the art.

[52] In one embodiment, the subject is or has been under cancer treatment. In one embodiment, the subject is or has been under a cancer treatment that is not an immune checkpoint inhibitor. In one embodiment, the subject is or has been under a cancer treatment that is an immune checkpoint inhibitor.

[53] In one embodiment, the subject is or has been treated by chemotherapy and/or radiotherapy, depending on the type of cancer.

[54] In one embodiment, the immune checkpoint inhibitor is selected from the group comprising an inhibitor of PD-1, an inhibitor of PD-L1, an inhibitor of CTLA-4 and a combination thereof.

[55] In one embodiment, the inhibitor of PD-1 is an anti-PD-1 antibody. In one embodiment, the inhibitor of PD-L1 is an anti-PD-L1 antibody. In one embodiment, the inhibitor of CTLA-4 is an anti-CTLA-4 antibody.

[56] Examples of PD-1 inhibitors include, without limitation, pembrolizumab and nivolumab.

[57] Examples of CLTA-4 inhibitors include ipilimumab.

[58] In one embodiment, the method previously described further comprises the step of comparing the fraction of AID-related mutations with a reference value.

[59] Typically, a reference value may be either implemented in the software or an overall median or other arithmetic mean across measurements may be built.

[60] In one embodiment, the reference value is obtained from a reference population.

[61] In one embodiment, the reference value is derived from the measurement of the fraction of the AID-related mutations according to the invention, in a reference population.

[62] In one embodiment, the reference value can be relative to a value derived from population studies defining a reference population, including without limitation, such subjects having similar age range, subjects in the same or similar ethnic group, similar cancer history and the like.

[63] In one embodiment, the reference value is derived from the measurement of the fraction of the AID-related mutations according to the invention, in a reference sample derived from one or more subject(s) in a reference population.

[64] In one embodiment, the reference population comprises subjects, preferably at least 50, more preferably at least 100, more preferably at least 200 subjects. In one embodiment, the reference population comprises at least 500 subjects.

[65] In one embodiment, the reference population comprises subjects having or having had a cancer, which are or have been treated with ICI, and which respond or have responded to the ICI.

[66] In one embodiment, the subject responds to the ICI if the administration of said ICI induces a positive response, *i.e.* a clinical benefit. Said positive response may a partial or a complete response. As used herein, a clinical benefit is a favorable effect on a meaningful aspect of how a subject feels (e.g., symptom relief), functions (e.g., improved mobility) or survives as a result of the treatment. Clinical benefit may be measured as an improvement or delay in the progression of a disease or condition.

[67] In one embodiment, the response to the ICI is evaluated by the RECIST (Response Evaluation Criteria In Solid Tumors) criteria. RECIST is a standard way to measure how well a cancer patient responds to treatment. It is based on whether tumors shrink, stay the same, or get bigger. To use RECIST, there must be at least one tumor that can be measured on x-rays, CT scans, or MRI scans. The types of response a patient can have are a complete response (CR), a partial response (PR), progressive disease (PD), and stable disease (SD).

[68] The iRECIST approach allows responses not typically observed in traditional systemic treatment to be identified and better documented. The guideline describes a standard approach to solid tumor measurement and definitions for objective change in tumor size which can be used in immunotherapy clinical trials. In addition, it defines the minimum amount of data to be collected in order to facilitate the creation of a data warehouse that can be used to later validate iRECIST.

[69] According to one embodiment, the predetermined fraction of AID-related mutation (*i.e.* the reference value) is obtained from a reference population as described hereinabove, wherein the fractions of AID-related mutations measured in a biological sample from each of the human subjects of the reference population are divided into

equal-sized groups by cut-off values referred to as “quantiles”, each group corresponding to a determined percentage of the fraction of AID-related mutations measured in the reference population. Examples of quantiles include, without being limited to, the median (defining 2 groups each comprising 50% of the fractions of AID-related mutations measured in the reference population), the terciles or tertiles (defining 3 groups each comprising a third of the fractions of AID-related mutations measured in the reference population), the quartiles (defining 4 groups each comprising 25% of the fractions of AID-related mutations measured in the reference population), the quintiles (defining 5 groups each comprising 20% of the fractions of AID-related mutations measured in the reference population) and the deciles (defining 10 groups each comprising 10% of the fractions of AID-related mutations measured in the reference population). In one embodiment, the “quantiles” are measured for each tumor type in a reference population

[70] According to the present invention, “quantiles” refer to the cut-off fraction of AID-related mutations below or above which lies a determined percentage of the fractions of AID-related mutations measured in the reference population. Therefore, the human subjects with a measured fraction of AID-related mutations below the first quantile are the human subjects with the lowest fractions of AID-related mutations, while the human subjects with a measured fraction of AID-related mutations above the last quantile are the human subjects with the highest fractions of AID-related mutations. For example, the 1st decile is the fraction of AID-related mutations below which 10% of the fractions of AID-related mutations measured in the reference population lie and above which 90% of the fractions of AID-related mutations measured in the reference population lie.

[71] Additionally, the term “quantiles” may also sometimes refer to the group so defined by said cut-off value. Thus, applied to the present invention, the term “quantiles” may also refer to the groups of the fractions of AID-related mutations measured in the reference population defined by the cut-off fraction of AID-related mutations. For example, the 1st decile may refer to the group of the fractions of AID-related mutations measured in the reference population corresponding to the lowest 10% of the fractions of AID-related mutations measured in the reference population. Accordingly, the 9th decile refers to the group of the fractions of AID-related mutations measured in the reference

population corresponding to the highest 10% of the fractions of AID-related mutations measured in the reference population. It follows that a fraction of AID-related mutations that is in the 1st decile is a fraction of AID-related mutations comprised in the lowest 10% of the fractions of AID-related mutations measured in the reference population and that a fraction of AID-related mutations that is in the 9th decile is a fraction of AID-related mutations comprised in the highest 10% of the fractions of AID-related mutations measured in the reference population.

[72] In one embodiment, the predetermined fraction of AID-related mutations is obtained from a reference population as described hereinabove, wherein the fractions of AID-related mutations measured in the reference population are divided into two equal-sized groups each corresponding to 50% of the fractions of AID-related mutations measured in the reference population.

[73] According to this embodiment of the present invention, the median of the fractions of AID-related mutations corresponds to the fraction of AID-related mutations below which 50% of the fractions of AID-related mutations measured in the reference population lie and above which 50% of the fractions of AID-related mutations measured in the reference population lie.

[74] Thus, in one embodiment, the predetermined fraction of AID-related mutations (*i.e* the reference value) is the median of the fractions of AID-related mutations of a reference population as described hereinabove. In one embodiment, the median is calculated for each tumor type.

[75] In one embodiment, the predetermined fraction of AID-related mutations is obtained from a reference population as described hereinabove, wherein the fractions of AID-related mutations measured in the reference population are divided into ten equal-sized groups each corresponding to 10% of the fractions of AID-related mutations measured in the reference population. As mentioned above, the cut-off values ("quantiles") so dividing the fraction of AID-related mutations measured in the reference population are called "deciles". Thus, in one embodiment, the predetermined fraction of AID-related mutations (*i.e.* the reference value) is a fraction of AID-related mutations

13

decile (*i.e.* first, second, third, fourth fifth, sixth, seventh, eighth or ninth decile) of a reference population as described hereinabove.

[76] According to this embodiment of the present invention:

- the fraction of AID-related mutations first decile corresponds to the fraction of AID-related mutations below which 10% of the fractions of AID-related mutations measured in the reference population lie and above which 90% of the fraction of AID-related mutations measured in the reference population lie;

- the fraction of AID-related mutations second decile corresponds to the fraction of AID-related mutations below which 20% of the fractions of AID-related mutations measured in the reference population lie and above which 80% of the fractions of AID-related mutations measured in the reference population lie;

- the fraction of AID-related mutations third decile corresponds to the fraction of AID-related mutations below which 30% of the fractions of AID-related mutations measured in the reference population lie and above which 70% of the fractions of AID-related mutations measured in the reference population lie;

- the fraction of AID-related mutations fourth decile corresponds to the fraction of AID-related mutations below which 40% of the fractions of AID-related mutations measured in the reference population lie and above which 60% of the fractions of AID-related mutations measured in the reference population lie;

- the fraction of AID-related mutations fifth decile corresponds to the fraction of AID-related mutations below which 50% of the fractions of AID-related mutations measured in the reference population lie and above which 50% of the fractions of AID-related mutations measured in the reference population lie;

- the fraction of AID-related mutations sixth decile corresponds to the fraction of AID-related mutations below which 60% of the fractions of AID-related mutations measured in the reference population lie and above which 40% of the fractions of AID-related mutations levels measured in the reference population lie;

- the fraction of AID-related mutations seventh decile corresponds to the fraction of AID-related mutations below which 70% of the fractions of AID-related mutations measured in the reference population lie and above which 30% of the fractions of AID-related mutations measured in the reference population lie;

- the fraction of AID-related mutations eight decile corresponds to the fraction of AID-related mutations below which 80% of the fractions of AID-related mutations measured in the reference population lie and above which 20% of the fractions of AID-related mutations measured in the reference population lie;

- the fraction of AID-related mutations ninth decile corresponds to the fraction of AID-related mutations below which 90% of the fractions of AID-related mutations measured in the reference population lie and above which 10% of the fractions of AID-related mutations measured in the reference population lie.

[77] In one embodiment, the predetermined fraction of AID-related mutations is obtained from a reference population as described hereinabove, wherein the fractions of AID-related mutations measured in the reference population are divided into three equal-sized groups each corresponding to a third of the fractions of AID-related mutations measured in the reference population. As mentioned above, the cut-off values ("quantiles") so dividing the fraction of AID-related mutations measured in the reference population are called "terciles" (or "tertiles"). Thus, in one embodiment, the predetermined fraction of AID-related mutations is a fraction of AID-related mutations tercile (or tertile) (*i.e.* first or second tercile) of a reference population as described hereinabove.

[78] In one embodiment, the predetermined fraction of AID-related mutations is obtained from a reference population as described hereinabove, wherein the fractions of AID-related mutations measured in the reference population are divided into four equal-sized groups each corresponding 25% of the fractions of AID-related mutations measured in the reference population. As mentioned above, the cut-off values ("quantiles") so dividing the fraction of AID-related mutations measured in the reference population are called "quartiles". Thus, in one embodiment, the predetermined fraction of AID-related

15

mutations is a fraction of AID-related mutations quartile (*i.e.* first, second or third quartile) of a reference population as described hereinabove.

[79] In one embodiment, the predetermined fraction of AID-related mutations is obtained from a reference population as described hereinabove, wherein the fractions of AID-related mutations measured in the reference population are divided into five equal-sized groups each corresponding to 20% of the fractions of AID-related mutations measured in the reference population. As mentioned above, the cut-off values ("quantiles") so dividing the fraction of AID-related mutations measured in the reference population are called "quintiles". Thus, in one embodiment, the predetermined fraction of AID-related mutations is a fraction of AID-related mutations quintile (*i.e.* first, second, third or fourth quintile) of a reference population as described hereinabove.

[80] In one embodiment, the subject is considered as being susceptible to respond to an ICI if the fraction of AID-related mutations of said subject is above the reference value as defined hereinabove.

[81] Thus, in one embodiment, the method as described hereinabove comprises the following steps:

a) assessing the fraction of AID-related mutations, wherein the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants,

b) comparing the fraction of AID-related mutations with a reference value,

wherein a fraction of AID-related mutations above the reference value is indicative of the subject with cancer as being susceptible to respond to a treatment with an immune checkpoint inhibitor.

[82] In one embodiment, the method as described hereinabove comprises the following steps:

a) assessing the fraction of AID-related mutations, wherein the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants,

b) comparing the fraction of AID-related mutations with a reference value, wherein the reference value is the median of the fractions of AID-related mutations measured in a reference population,

wherein a fraction of AID-related mutations above the reference value is indicative of the subject with cancer as being susceptible to respond to a treatment with an immune checkpoint inhibitor.

[83] In one embodiment, the method as described hereinabove further comprises treating the subject identified as being susceptible to respond to a treatment with ICI, by administering a therapeutically effective amount of an ICI.

[84] Thus, in one aspect, the invention further relates to an *in vitro* method for identifying a subject with cancer as being susceptible to respond to a treatment with an immune checkpoint inhibitor, the method comprising:

a) assessing the fraction of AID-related mutations, wherein the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants,

b) comparing the fraction of AID-related mutations with a reference value,

c) identifying a subject with cancer as being susceptible to respond to a treatment with ICI when the fraction of AID-related mutations is above the reference value as assessed in b), and

d) treating the subject being susceptible to respond to a treatment with ICI identified at step c), by administering a therapeutically effective amount of an ICI.

[85] In one aspect, the invention also relates to the use of the AID-related mutations, in particular the fraction of AID-related mutations, as a biomarker for identifying a subject with cancer as being susceptible to respond to a treatment with an ICI.

[86] In one aspect, the invention relates to a method for preventing and/or treating cancer in a subject identified by the method as described hereinabove, comprising the administration of a therapeutically efficient amount of an ICI.

[87] In one embodiment, the ICI is to be administered at a therapeutically effective amount to the subject identified by the method as described hereinabove.

[88] In one embodiment, the therapeutically effective amount is defined according to the treatment of the subject, the type of cancer and the ICI used.

[89] In one embodiment, the ICI is formulated as a pharmaceutical composition.

[90] Within the meaning of the invention, the expression “pharmaceutical composition” refers to a composition comprising an active principle in association with a pharmaceutically acceptable vehicle or excipient. A pharmaceutical composition is for therapeutic use, and relates to health.

[91] Within the meaning of the invention, the expression “pharmaceutically acceptable excipient" refers to an inert vehicle or carrier used as a solvent or diluent in which the pharmaceutically active agent is formulated and/or administered, and which does not produce an adverse, allergic or other reaction when administered to an animal, preferably a human. This includes all solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonic agents, absorption retardants and the like. For human administration, preparations must meet standards of sterility, general safety and purity as required by regulatory agencies, such as the FDA or EMA.

[92] In one embodiments, the ICI, or the pharmaceutical composition may be administered to a subject in need thereof by any suitable route, *i.e.*, by an oral administration, a topical administration or a parenteral administration, *e.g.*, by injection, including a sub-cutaneous administration, a venous administration, an arterial

administration, in intra-muscular administration, an intra-ocular administration and an intra-auricular administration.

[93] In one embodiment, the method as described hereinabove is for prognosing survival of a subject with cancer and being treated with ICI.

[94] In one embodiment, the cancer is a solid cancer. In one embodiment, the cancer is a blood cancer.

[95] In one embodiment, the cancer is as described previously. In one embodiment, the cancer is selected form the group comprising melanoma, non-small-cell lung carcinoma (NSCLC), renal cell carcinoma, head and neck cancers, merkel-cell carcinoma, gastric cancer, small-cell lung carcinoma (SCLC), Hodgkin lymphoma, breast cancer, cervical cancer, colorectal cancer, endometrial cancer, hepatocellular cancer, esophageal cancer, mesothelioma, MSI (microsatellite instability)-high solid tumors, TMB (tumor mutation burden)-high tumors, breast cancer and urothelial carcinoma.

[96] In one embodiment, the subject suffers from a cancer.

[97] In one embodiment, the subject has been previously diagnosed with cancer by authorized personnel skilled in the art.

[98] In one embodiment, the subject is or has been under cancer treatment. In one embodiment, the subject is or has been treated with chemotherapy and/or radiotherapy.

[99] In one embodiment, the subject is or has been treated with ICI as described hereinabove. In one embodiment, the ICI is selected from the group comprising an inhibitor of PD-1, an inhibitor of PD-L1, an inhibitor of CTLA-4 and a combination thereof.

[100] In one embodiment, the method previously described further comprises the step of comparing the fraction of AID-related mutations with a reference value.

[101] Typically, a reference value may be either implemented in the software or an overall median or other arithmetic mean across measurements may be built.

[102] In one embodiment, the reference value is obtained from a reference population.

[103] In one embodiment, the reference value is derived from the measurement of the fraction of the AID-related mutations according to the invention, in a reference population.

[104] In one embodiment, the reference value can be relative to a value derived from population studies, including without limitation, such subjects having similar age range, subjects in the same or similar ethnic group, similar cancer history and the like.

[105] In one embodiment, the reference value is derived from the measurement of the fraction of the AID-related mutations according to the invention, in a reference sample derived from one or more subject(s) in a reference population.

[106] In one embodiment, the reference population comprises subjects, preferably at least 50, more preferably at least 100, more preferably at least 200 and even more preferably at least 500 subjects.

[107] In one embodiment, the reference population comprises subjects having (or suffering) or having had cancer, which are or have been treated with ICI, and which respond or have responded to the ICI.

[108] In one embodiment, the predetermined fraction of AID-related mutation (*i.e.* the reference value) is measured as described previously.

[109] In one embodiment, the predetermined fraction of AID-related mutations (*i.e.* the reference value) is the median of the fractions of AID-related mutations of a reference population as described hereinabove. In one embodiment, the median is calculated for each tumor type.

[110] In one embodiment, the predetermined fraction of AID-related mutations (*i.e.* the reference value) is a fraction of AID-related mutations decile (*i.e.* first, second, third, fourth, fifth, sixth, seventh, eighth or ninth decile) of a reference population as described hereinabove.

20

[111] In one embodiment, a fraction of AID-related mutations above the reference value is associated with a high prognosis (*i.e.* after receiving ICI).

[112] In one embodiment, a fraction of AID-related mutations below the reference value is associated with a low prognosis (*i.e.* after receiving ICI).

[113] Thus, in one embodiment, the method as described hereinabove comprises the following steps:

a) assessing the fraction of AID-related mutations, wherein the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants,

b) comparing the fraction of AID-related mutations with a reference value,

wherein a fraction of AID-related mutations above the reference value is indicative of the subject with cancer and being treated with ICI as having a high prognosis (*i.e.* after receiving ICI).

[114] In one embodiment, the method as described hereinabove comprises the following steps:

a) assessing the fraction of AID-related mutations, wherein the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants,

b) comparing the fraction of AID-related mutations with a reference value, wherein preferably the reference value is the median of the fractions of AID-related mutations measured in a reference population,

wherein a fraction of AID-related mutations above the reference value is indicative of the subject with cancer and being treated with ICI as having a high prognosis (*i.e.* after receiving ICI).

21

[115] In one embodiment, the subject identified by the method of prognosis survival as described hereinabove is treated with an adapted care. In one embodiment, the adapted care is a cancer treatment adapted depending on the prognosis of the subject.

[116] In one embodiment, the subject associated with a low prognosis is treated with aggressive cancer treatments such as radiations, chemotherapies.

[117] In one embodiment, the subject associated with a high prognosis is treated with less aggressive cancer treatments.

[118] In one aspect, the present invention further relates to the use of AID-related mutations, preferably the fraction of AID-related mutations, in an *in vitro* method for prognosing survival of a subject with cancer and being treated with ICI as described hereinabove.

[119] The present invention also relates to a computer system for prognosing survival in a subject affected with a cancer and being treated with ICI, using AID-related mutations, in particular the fraction of AID-related mutations, as described hereinabove. The present invention also related to a computer-implemented method for prognosing survival in a subject, using AID-related mutations, in particular the fraction of AID-related mutations, as described hereinabove.

[120] The present invention also relates to a computer system for determining a personalized course of treatment in a subject affected with a cancer and being treated with ICI, using AID-related mutations, in particular the fraction of AID-related mutations, as described hereinabove. The present invention also relates to a computer-implemented method for determining a personalized course of treatment in a subject affected with a cancer and being treated with ICI, using AID-related mutations, in particular the fraction of AID-related mutations, as described hereinabove.

[121] As used herein, the term “computer system” refers to any and all devices capable of storing and processing information and/or capable of using the stored information to control the behavior or execution of the device itself, regardless of whether such devices are electronic, mechanical, logical, or virtual in nature. The term “computer system” can

refer to a single computer, but also to a plurality of computers working together to perform the function described as being performed on or by a computer system. A method implemented using a computer system is referred to as a "computer-implemented method".

[122] In one embodiment, the computer system according to the present invention comprises:

- at least one processor, and
- at least one computer-readable storage medium that stores code readable by the processor.

[123] As used herein, the term "processor" is meant to include any integrated circuit or other electronic device capable of performing an operation on at least one instruction word, such as, e.g., executing instructions, codes, computer programs, and scripts which it accesses from a storage medium. However, the term "processor" should not be construed to be restricted to hardware capable of executing software, and refers in a general way to a processing device, which can for example include a computer, a microprocessor, an integrated circuit, or a programmable logic device (PLD). The processor may also encompass one or more graphics processing units (GPU), whether exploited for computer graphics and image processing or other functions. Additionally, the instructions and/or data enabling to perform associated and/or resulting functionalities may be stored on any processor-readable medium, including, but not limited to, an integrated circuit, a hard disk, a magnetic tape (including floppy disk and zip diskette), an optical disc (including Blu-ray, compact disc and digital versatile disc), a flash memory (including memory card and USB flash drive) a random-access memory (RAM) (including dynamic and static RAM), a read-only memory (ROM) or a cache. Instructions may be in particular stored in hardware, software, firmware or in any combination thereof.

[124] Examples of processors include, but are not limited to, central processing units (CPU), microprocessors, digital signal processors (DSPs), general purpose microprocessors, application specific integrated circuits (ASICs), field programmable logic arrays (FPGAs), and other equivalent integrated or discrete logic circuitry.

[125] The present invention also related to a computer program comprising software code readable by the processor adapted to perform, when executed by said processor, the computer-implemented methods as described herein.

[126] The present invention also relates to a computer-readable storage medium comprising code readable by the processor which, when executed by said processor, causes the processor to carry out the steps of the computer-implemented methods as described herein.

[127] Examples of computer-readable storage medium include, but are not limited to, an integrated circuit, a hard disk, a magnetic tape (including floppy disk and zip diskette), an optical disc (including Blu-ray, compact disc and digital versatile disc), a flash memory (including memory card and USB flash drive) a random-access memory (RAM) (including dynamic and static RAM), a read-only memory (ROM) or a cache.

[128] In one embodiment, the computer-readable storage medium is a non-transitory computer-readable storage medium.

[129] In one embodiment, the code stored on the computer-readable storage medium, when executed by the processor of the computer system, causes the processor to:

- receive an input level, *i.e.* the fraction of AID-related mutations determined in a sample previously obtained from the subject,
- analyze and transform the input level by organizing and/or modifying each input level to derive a probability score and/or a classification label via at least one machine learning algorithm,
- generate an output, wherein the output is the classification label and/or the probability score, and
- provide a prognosis of survival of the subject based on the output; or
- provide a personalized course or information to determine a personalized course of treatment for the subject based on the output.

[130] In one embodiment, the code stored on the computer-readable storage medium, when executed by the processor of the computer system, causes the processor to:

- receive an input level, the fraction of AID-related mutations determined in a sample previously obtained from the subject,
- analyze and transform the input level by organizing and/or modifying each input level to derive a probability score and/or a classification label via at least one machine learning algorithm,
- generate an output, wherein the output is the classification label and/or the probability score, and
- provide a prognosis of survival of the subject based on the output; or
- provide a personalized course or information to determine a personalized course of treatment for the subject based on the output.

[131] As used herein, the terms “learning algorithm” or “machine learning algorithm” refer to computer-executed algorithms that automate analytical model building, e.g., for clustering, classification or profile recognition. Learning algorithms perform analyses on training datasets provided to the algorithm. Learning algorithms output a “model”, also referred to as a “classifier”, “classification algorithm” or “diagnostic algorithm”. Models receive, as input, test data and produce, as output, an inference or a classification of the input data as belonging to one or another class, cluster group or position on a scale, such as diagnosis, stage, prognosis, disease progression, responsiveness to a drug, etc.

[132] “Datasets” are collections of data used to build a machine learning mathematical model, so as to make data-driven predictions or decisions. In “supervised learning” (i.e., inferring functions from known input-output examples in the form of labelled training data), three types of machine learning datasets are typically dedicated to three respective kinds of tasks: “training”, i.e., fitting the parameters; “validation”, i.e., tuning machine learning hyperparameters (which are parameters used to control the learning process);

and “testing”, i.e., checking independently of a training dataset exploited for building a mathematical model that the latter model provides satisfying results.

[133] A variety of learning algorithms can be used to infer a condition or state of a subject. Machine learning algorithms may be supervised or unsupervised. Learning algorithms include, but are not limited to, artificial neural networks (e.g., back propagation networks), discriminant analyses (e.g., Bayesian classifier, Fischer analysis), support vector machines, decision trees (e.g., recursive partitioning processes, such as classification and regression trees [CART]), random forests, linear classifiers (e.g., multiple linear regression [MLR], partial least squares [PLS] regression, principal components regression [PCR]), hierarchical clustering and cluster analysis. The learning algorithm generates a model or classifier that can be used to make an inference, e.g., an inference about a disease state of a subject.

[134] In one embodiment, the at least one machine learning algorithm was previously trained with at least one training dataset.

[135] In one embodiment, the at least one training dataset comprises information relating to the level, the fraction of AID-related mutations from samples previously obtained from reference subjects.

[136] In one embodiment, the at least one training dataset comprises information relating to the level, the fraction of AID-related mutations from samples previously obtained from a reference population having cancer.

[137] In one embodiment, the at least one machine learning algorithm is selected from the group comprising an artificial neural network (ANN), a perceptron algorithm, a deep neural network, a clustering algorithm, a k-nearest neighbors algorithm (k-NN), a decision tree algorithm, a random forest algorithm, a linear regression algorithm, a logistic regression algorithm, a linear discriminant analysis (LDA) algorithm, a quadratic discriminant analysis (QDA) algorithm, a support vector machine (SVM), a Bayes algorithm, a simple rule algorithm, a clustering algorithm, a meta-classifier algorithm, a Gaussian mixture model (GMM) algorithm, a nearest centroid algorithm, a gradient boosting algorithm (such as, e.g., an extreme gradient boosting [XG Boost] algorithm or

an adaptative boosting [AdaBoost] algorithm), a linear mixed effects model algorithm, and a combination thereof.

BRIEF DESCRIPTION OF THE DRAWINGS

[138] **Figure 1** is a forest plot from Miao et al. study showing the overall impact estimated using multivariate Cox proportional hazard model of the fraction of AID mutations after adjustment by TMB (top 20%), age and gender.

[139] **Figure 2** is a combination of three forest plots showing the meta-analysis of the survival impact of the fraction of AID mutations in different studies. **Figure 2A** represent the effect of using AID/APOBEC (5th decile as cut-off) or SNV substitutions where AID remains significant across all the studies. Assessment of the prognostic value of the fraction of AID mutations in the IMPACT study. Symbols are filled if reaching significance or empty otherwise. **Figure 2B** represents a forest plot of a Cox model of the global impact, after adjustment by TMB (top 20%), median APOBEC mutations, age and gender. **Figure 2C** represents a forest plot of the Cox model of the impact of AID mutations per cancer subtype.

[140] **Figure 3** is a combination of a Kaplan-Meier plot, a forest plot and two histograms showing the impact of AID-related mutations in metastatic melanoma from Liu et al. study. **Figure 3A** is a Kaplan-Meier plot showing a better overall survival in patients with higher fraction of AID mutations (according to the median). **Figure 3B** shows the distribution using boxplots of the fraction of AID mutations according to the best response under ICI treatment, PD (progression disease), PR/CR (partial response/complete response) and SD/MR (stable disease/mixed response). **Figure 3C** shows the distribution using boxplots of the fraction of AID mutations according to the localization of the melanoma. **Figure 3D** is a forest plot of the Cox proportional hazards ratio multivariate model adjusting the potential prognostic impact on overall survival (OS) of the fraction of AIDA mutations (according to the median) with tumor purity and gender.

27

[141] **Figure 4** is a combination of a forest plot and one histogram showing the impact of AID-related mutations from Pender et al. study. **Figure 4A** is a forest plot of the Cox proportional hazard ratio multivariate model to assess the prognostic impact on OS of the fraction of AID mutations (according to the median), adjusted by TMB, age and gender. **Figure 4B** is the distribution of the fraction of AID mutations according to the clinical benefit of using ICI, defined using NCB (no clinical benefit) and DCB (durable clinical benefit).

[142] **Figure 5** is a combination of a Kaplan-Meier plot and a forest plot showing the impact of AID-related mutations from Hugo et al. study. **Figure 5A** is a Kaplan-Meier plot comparing the OS according to the fraction of AID mutations (cut-off median). **Figure 5B** is a forest plot of the Cox proportional hazard ratio multivariate model to assess the prognostic value of the fraction of AICDA mutations on OS adjusting by TMB (cut-off ≥ 10).

[143] **Figure 6** is a combination of a Kaplan-Meier plot and a forest plot showing the impact of AID-related mutations from the Braun et al. study. **Figure 6A** is a Kaplan-Meier plot comparing the OS according to the fraction of AID mutations (cut-off media). **Figure 6B** shows a forest plot of the Cox proportional hazard ratio multivariate model to assess the prognostic value of the fraction of AID mutations on OS adjusting by gender, age (cut-off median) and PBRM1 mutations.

[144] **Figure 7** is a forest plot showing the meta-analysis of the survival impact of the fraction of AID mutations in different studies. It represents the effect using all the deciles of the fractions of AID mutations (adjusting every decile of fractions of AID mutations per TMB $\geq$ 10 mut / Mb), the overall impact of AID with a better OS is present independently of the cut-off. Symbols are filled if reaching significance or empty otherwise.

[145] **Figure 8** is a forest plot showing the meta-analysis of the survival impact of the fraction of AID mutations in different studies. It represents the effect using all the deciles of the fractions of AID mutations at univariate level, the overall impact of AID with a

better OS is present independently of the cut-off. Symbols are filled if reaching significance or empty otherwise.

[146] **Figure 9** is a combination of two forest plots showing the assessment of the prognostic value of the fraction of AID mutations in the IMPACT study. It shows the forest plot of a Cox model of either the global impact, after adjustment by TMB (≥10 mut/Mb), median APOBEC mutations, age and gender (**Figure 9A**) or the impact by tumor subtype (**Figure 9B**).

EXAMPLES

[147] The present invention is further illustrated by the following examples.

Materials and Methods

Subject details

[148] The total cohort consisted of 50,631 tumor samples representing more than 80 cancer types. TCGA information consisted of: mutational data in Mutation Annotation Format (MAF) included 9,264 cancer patients (24 cancer types) and 741 normal samples; RNA-seq data (read counts, n = 9,101); immune data (i.e. cibersort calculated immune populations, n = 8,983), allele-specific integer copy numbers (ABSOLUTE calculated, n = 7,216), available viral counts (n = 5,741) and previously predicted neo-epitopes (n = 2,143; cancer types = 14). The PCAWG data (ICGC) included 2,775 cancer patients along 35 different cancer types with WGS information (SNV and CNV) from which 1,522 had the expression data available. Composite mutations data included 31,353 cancer patients from the MSKCC comprising 41 tumor types by the MSK-IMPACT assay (sizes depending on the date of sequencing comprising 341, 410, and 468 cancer-associated targeted genes) downloaded from CBioPortal for the general maf or their github repository (https://github.com/taylor-lab/composite-mutations/tree/master/data) for the clinical, mutational burden classification, mutational signatures, composite mutation annotation, phasing information and molecular timing (Gorelick et al., Phase and context shape the function of composite oncogenic mutations, 2020, Nature 582, 100–103).

29

Additionally, hematological cancers cohort (AML, DLBCL, Myelodysplastic Syndromes and other leukemias; n = 3,859) (for The St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project et al., The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias, 2015, Nat. Genet. 47, 330–337; Holmfeldt et al., The genomic landscape of hypodiploid acute lymphoblastic leukemia, 2013, Nat. Genet. 45, 242–252; Landau et al., Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia, 2013, Cell 152, 714–726; Landau et al., Mutations driving CLL and their evolution in progression and relapse., 2015, Nature 526, 525–530; Lohr et al., Widespread Genetic Heterogeneity in Multiple Myeloma: Implications for Targeted Therapy, 2014, Cancer Cell 25, 91–101; Nangalia et al., Somatic CALR Mutations in Myeloproliferative Neoplasms with Nonmutated JAK2, 2013, N. Engl. J. Med. 369, 2391–2405; Papaemmanuil et al., Genomic Classification and Prognosis in Acute Myeloid Leukemia, 2016, N. Engl. J. Med. 374, 2209–2221; Puente et al., Non-coding recurrent mutations in chronic lymphocytic leukaemia, 2015, Nature 526, 519–524; Quesada et al., Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia, 2012, Nat. Genet. 44, 47–52; Reddy et al., Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma, 2017, Cell 171, 481-494.e15; the St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project et al., Deregulation of DUX4 and ERG in acute lymphoblastic leukemia, 2016 Nat. Genet. 48, 1481–1489; Tyner et al., Functional genomic landscape of acute myeloid leukaemia, 2018, Nature 562, 526–531; Welch et al., TP53 and Decitabine in Acute Myeloid Leukemia and Myelodysplastic Syndromes, N. Engl. J. Med. 375, 2023–2036; Yoshida et al., Frequent pathway mutations of splicing machinery in myelodysplasia, 2011, Nature 478, 64–69) and pediatric cancers cohort (20 tumor types; n = 1,051) (Gröbner et al., The landscape of genomic alterations across childhood cancers, 2018, Nature 555, 321–327) were used. ICI cohort consisted of 2,261 samples coming from: MSKCC-IMPACT dataset (n = 1,472; 11 tumor types), Pender et al. cohort (n = 98, 19 tumor types), Miao et al. cohort (n = 249, four six tumor types), Liu et al. (n = 144, melanoma), Hugo et al (n = 37, melanoma) and Braun et al (n = 261; ccRCC) (Braun et al., Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma, 2020, Nat. Med. 26, 909–918; Hugo et al., Genomic and

Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma, 2016, Cell 165, 35–44; Liu et al., Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma, 2019, Nat. Med. 25, 1916–1927; Miao et al., Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors, 2018, Nat. Genet. 50, 1271–1281; Pender et al., Genome and Transcriptome Biomarkers of Response to Immune Checkpoint Inhibitors in Advanced Solid Tumors, 2021, Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 27, 202–212; Samstein et al., Tumor mutational load predicts survival after immunotherapy across multiple cancer types, 2019, Nat. Genet. 51, 202–206). Riaz et al. melanoma cohort consisted of 68 patients treated with Nivolumab (anti-PD-1) from which 35 had previously progressed on Ipilimumab (anti-CTLA-4) treatment from which data was obtained prior treatment (pre) or 4 weeks after initiation of Nivo (on). Data consisted of WES, neo-epitopes (npre = 68; non = 41) and RNA-seq (npre = 45; non = 41) (Riaz et al., Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab, 2017, Cell 171, 934-949.e16).

Tracking AICDA-related mutations

[149] A code was develop to detect *AICDA*-related mutations over *wrCy/rGyw* (+/- strand, where "W" stands to either adenine or thymine, "R" to purine and "Y" to pyrimidine) motifs, giving a total of 8 motifs per strand (positive strand = AACC, AACT, AGCC, AGCT, TACC, TACT, TGCC, TGCT; negative strand = TTGG, TTGA, TCGG, TCGA, ATGG, ATGA, ACGG, ACGA), and its enrichment around a 60 bp flanking sequences (the code allows other bp windows for the user). The enrichment strength over the wrCy/rGyw motifs was calculated as:

$$EAICDA = \frac{Mutations\ wrcy\ X\ Context\ C\ or\ G}{Mutations\ C\ or\ G\ X\ Context\ wrcy}$$

[150] Then a Fisher's exact test was applied to evaluate the over-representation of AID mutations in each samples by comparing the ratio of substitutions in and out of the AID preferred motifs to the ratio of all cytosines and guanines occurring within the provided genome window around the mutation (60 bp by default), similar to what was previously described for APOBEC (Roberts et al., An APOBEC cytidine deaminase mutagenesis

pattern is widespread in human cancers, 2013, Nat. Genet. 45, 970–976). The code was developed under R, takes a maf (mutation annotation format) object as input and outputs a S3 class object containing: i) a matrix of the 768 possible tetranucleotide substitutions across the samples; ii) a data table with all the needed values for enrichment calculation, the enrichment score, Fisher exact-test p-value and fdr for enrichment, fraction of AID mutations, among others; and iii) a maf like data table, with the same format as the input, containing only the attributed AID mutations. Finally, mutations were tagged as AID or not AID if overlapping or not with the mutations found in the output maf after applying the function.

AICDA motifs distribution across the genome

[151] The *AICDA* motifs, *i.e.* WRCY motifs W=adenine or thymine, R=purine, C=cytosine, Y=pyrimidine, that is: AACC, AGCC, AACT, AGCT, TACC, TGCC, TACT and TGCT. A R script was used to find these patterns in the GRCh38 genome using the Biostrings package v2.60.0. In addition, the number of these AICDA motifs was also calculated in 20kb binned windows throughout the genome using bedtools, adjusting by the chromosome size.

Attributing mutations to mutagenic processes

[152] Mutations previously tagged as not AICDA were subjected to signature attribution to 46 (signatures SBS: 27, 39, 43, 45-60 were excluded since they are attributed to sequencing artefacts) of the 65 COSMIC mutational signatures (v3.0) using Palimpsest package with default parameters (Alexandrov et al., Signatures of mutational processes in human cancer, 2013, Nature 500, 415–421; Alexandrov et al., The repertoire of mutational signatures in human cancer, 2020, Nature 578, 94–101; Shinde et al., Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer, 2018, Bioinformatics). To avoid over-fitting, signatures not contributing with at least one mutation within 50% of the samples per tumor type (Median-SBSnTumorX < 1) were removed and mutations were re-fitted using the remaining signatures. Furthermore, signatures proportions per sample were re-calculated adding the number of previously identified AICDA mutations to the signature data.

Signatures were not calculated from the MSKCC-Composites, and ICI cohorts because it was already available or not used.

Attributing mutations as driver genes

[153] Positively-selected genes per tumor type within each cohort were obtained by calculating dN/dS likelihood ratios (dNdScv package) through negative binomial regression modeling of the background mutation rate of each gene using distinct genomic covariates including variation in mutation density across genes, context-dependent substitutions (mutational signatures), transcriptional strand bias, chromatin state, expression and replication time. Additionally, ultra-hypermutator samples and extremely mutated genes per sample were removed, to avoid loss of sensitivity. Genes were considered as drivers if having q-values < 0.01 (Benjamini-Hodgberg's multiple testing correction of p-values) (Martincorena et al., Universal Patterns of Selection in Cancer and Somatic Tissues, 2017, Cell 171, 1029-1041.e21). In addition, in the ICGC cohort, the selection intensity of every particular mutation related with APOBEC or AICDA signatures by deconvolution of prevalence by mutation rates was used for recurrent amino acid mutations within three oncoproteins caused by single-nucleotide changes using cancereffectsizeR v2.1.3 package. The observed substitution rates were divided by the expected substitution rates in the absence of selection. The expected substitution rates in the absence of selection were calculated as the average per-site synonymous mutation rate of the gene, normalized for the average weight of trinucleotide mutational signature burden for that signature. The quotient of observed to expected numbers of substitutions was the selection intensity, as previously described (Cannataro et al., Effect Sizes of Somatic Mutations in Cancer, 2018, JNCI J. Natl. Cancer Inst. 110, 1171–1177).

Statistical analyses and figures

[154] All statistical analyses were performed using the R statistical programming environment (version 4.0). Figures were generated using either base R or the ggplot2 library. Mann Whitney U test was used for differences in distributions between two population groups, unless otherwise noted. Overall survival analysis to ICI was assessed using log-rank Kaplan-Meier curves and univariate/multivariate Cox proportional

hazards regression modeling. Several Cox proportional models were assessed for every study (i.e. analyzing the deciles, from 10th to 90th, of the fraction of AID induced mutations in every included study, unadjusted, using the median of the fraction of AID mutations, and also these models were adjusted by TMB >= 10mut/Mb). To combine the different survival models, a random-effects model was used with the meta v4.18-1 package (Schwarzer et al., Fixed Effect and Random Effects Meta-Analysis, 2015, In Meta-Analysis with R, G. Schwarzer, J.R. Carpenter, and G. Rücker, eds. (Cham: Springer International Publishing), pp. 21–53), using log hazard ratio and standard errors of each model per study. The inverse variance method was used for pooling. The random-effects estimate was based on the DerSimonian-Laird method (DerSimonian and Laird, Meta-Analysis in Clinical Trials Revisited, 2015, Contemp. Clin. Trials 45, 139–145). The meta-analysis results were represented in a forestplot using the forestplot function of the ggforestplot v0.1.0 package.

Results

Landscape of AID-related mutations at pan-cancer level

[155] AID-related mutations were found in the vast majority of cancers studied. Overall, the AID-related mutations were found in roughly 5.2% (5.1-5.3% at 95% confidence interval[CI]) and 6.6% (6.5-6.8% at 95% CI) in APOBEC mutations. When the AID-related to the Single-Base Substitution (SBS) somatic signature according to Alexandrov (Alexandrov et al., The repertoire of mutational signatures in human cancer, 2020, Nature 578, 94–101) were included, similar results were found at pan-cancer level using the WGS ICGC dataset. Interestingly, the distribution of the somatic signature related to AID was homogeneously identified in the vast majority of cancers. Likewise, a similar frequency was found in the pan-cancer TCGA, MSKCC cohort and different pediatric dataset. Conversely, as expected, the frequency of AID-related mutations was slightly higher in hematological cancers at approximately 8%. Additionally, in regard to the genomic distribution of *AICDA* motifs (in the normal genome) and mutations (within tumors), for the majority of tumor types the highest density of mutations were located in chromosome 5, in which *GPR98* and *DNAH5* were frequently affected, followed by chromosome 17 and 2, though the number of *AICDA* motifs in the genome is higher in the latest (0.01 and 3.2x10-4, respectively; FDR corrected p-value Wilcoxon-test) but due

to its chromosome length. In diffuse large B-cell lymphoma (DLBCL), most commonly affected chromosomes involved the presence of either immunoglobulin related genes: *IGH* (chr14), *IGL* (chr22), *IGK* (chr2) or genes already related with off-target AID activity: *PIM1, IRF4, HIST1H1C* (chr6). Interestingly, within the driver genes context hematological cancers (i.e. Lymph-BNHL, DLBCL) and medulloblastoma had the highest signature contribution of AID provoked mutations. Furthermore, among the involved targets, *TP53*, in all cohorts; *IDH1*, in hematological cancers, *GBM* and *LGG*; and *PIK3* genes (TCGA and ICGC cohorts), were recurrently altered. These results were also confirmed using a selection intensity approach of every somatic in the ICGC dataset, showing a higher selection intensity *of PIK3CA, NFE2L2* but also in "minor" *IDH1* mutations (i.e. not *R132H*) and *PTEN*.

[156] The AID signature was more frequently negatively correlated with the tumor mutation burden (TMB) of cancers from TCGA (*i.e*. in adenoid cystic carcinoma (ACC), kidney renal papillary cell carcinoma (KIRP), kidney renal clear cell carcinoma (KIRC), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), ovarian cancer (OV) and thyroid cancer (THCA)). Conversely, only in bladder carcinoma (BLCA) the AID signature was slightly positively correlated with TMB, p=0.01, R=0.13. Furthermore, *AICDA* expression was not associated with age, and was slightly negatively associated in colon adenocarcinoma (COAD) (Rho=-0.18, p=7.1e-05) and KIRP (Rho=-0.12, p=0.04).

[157] AID mutations were found in younger patients when compared to APOBEC mutations (median age 61 years vs 65 years p = 0.009). However, within AID mutations, no difference of age was found according to gender, but only in the APOBEC related mutations where these mutations were found in elderly men compared to women ($p = 7.3x10^{-7}$, Wilcoxon-test).

The impact of AID-related mutations with immune checkpoint inhibitor (ICI) response

[158] Different available datasets analyzed by different sequencing approaches were used: WGS, a pan-cancer dataset with 19 different types of cancers (n=98) using virtually all types of ICI or combinations (i.e. anti-PD-1, anti-PD-L1, anti-CTLA-4, anti-PD-1 +

CTLA-4, anti-PD-L1 + CTLA-4, among others) (Pender et al., Genome and Transcriptome Biomarkers of Response to Immune Checkpoint Inhibitors in Advanced Solid Tumors, 2021, Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 27, 202–212), WES in 6 types of MSS cancers (n=249) treated with anti-PD-1, anti-PD-L1, anti-CTLA-4, or a combination of these therapies (Miao et al., Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors, 2018, Nat. Genet. 50, 1271–1281), WES in metastatic melanoma (n=144) treated with anti-PD-1 (Liu et al., Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma, 2019, Nat. Med. 25, 1916–1927), targeted next generation sequencing (NGS), IMPACT-MSKCC, of 10 different cancer types (n=1472), WES in metastatic melanoma (n=37) treated with anti-PD-1 (Hugo et al., Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma, 2016, Cell 165, 35–44) and WES in clear cell renal cell carcinoma (ccRCC) (n=261) treated with anti-PD-1 (Braun et al., Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma, 2020, Nat. Med. 26, 909–918), with overall more than 2000 patients analyzed with sequencing data treated with ICI. A random-effects meta-analysis comparing the overall survival (OS) of all these studies was performed and compared the impact of AID, to the APOBEC signature and the different single nucleotide variants (SNV), Figure 2. The details of this analysis are provided in the methods. Strikingly, the AID signature was associated with the best OS in all of the studies and the random-effects model showed also a favorable prognosis (median as the cut-off), Figure 2A. Moreover, the effect was still significant across almost all the studies independently of either decile chosen as cut-off at univariate (Figure 8) or multivariate adjusting for TMB (Figure 7). Accordingly, the APOBEC signature was associated with a favorable prognosis, but not in all datasets. However, the random-effects model also indicated an overall favorable prognosis associated with APOBEC, Figure 2A. The rest of SNV showed much more heterogeneous results and only T>A and T>G mutations were associated with favorable prognosis in the random-effects model, Figure 2A.

[159] Interestingly, within the largest study of IMPACT-MSKCC, the fraction of AID-related mutations (using the top 50% of all histologies as a cut-off) was also

independently associated with both better OS (Hazard ratio (HR) 0.715 [95% CI 0.61−0.839] with a p=3.81e$^{10-5}$) and predictive value compared to TMB or APOBEC, after adjusting by TMB (top 20% of each histology as the cut-off), APOBEC signature (top 50% of all histologies as the cut-off) age and sex, Figures 2B. It should be noted, that using an univariate Cox proportional Hazards ratio model per every cancer type or adjusting TMB ≥10, the results were also similar in the overall population of this study. In addition, the clinical impact of AID-related fraction of mutations was found in metastatic melanoma and cancer with unknown primary (Figure 2C, Figures 9A-B). Additionally, there was practically no correlation between the fraction of AID mutations with the APOBEC signature neither globally nor by tumor type in this cohort and in the ICGC and TCGA datasets (Spearman correlation).

[160] In the study of Miao et al. (Miao et al., 2018), the fraction of AID-related mutations was associated with an improved OS using the overall population, HR=0.241 [95% CI 0.126−0.46] and p=3.81e$_{10-5}$, after adjusting by age, sex and TMB (Figure 1). Likewise, in the study of Liu et al. with metastatic melanoma, the presence of high fraction of AID-related mutations was also associated with a better OS, in univariate analysis using Kaplan-Meier plot with log-rank p=0.0025, but also in a multivariate Cox model, HR=0.5 [95% CI 0.31-0.8] p=0.004, after adjusting by tumor purity and gender (Figures 3A;D). Noteworthy, when the response in this study was stratified according to the best response under anti-PD-1, the patients showing a response (either partial response, PR or complete response, CR) had a significantly higher fraction of AID-related mutations compared to the patients with progressive disease, PD or with patients with stable disease, SD, Figure 3B. Interestingly, the localization of the melanoma (i.e. acral, mucosal, occult and skin) was stratified and the only subtype with significant higher level of fraction of AID-related mutations in responders versus non-responders was the skin, Figure 3C. Furthermore, in the Pender et al study (Pender et al., 2021) using WGS of a pan-cancer dataset, the results were also in the same line with a statistically significant effect of the fraction of AID mutations with better OS, HR=0.62 [95% CI 0.38-0.99], p=0.048, in the multivariate Cox model, after adjusting by TMB, gender and age, Figure 4A. Importantly, the clinical benefit (durable clinical benefit, DCB, versus non-durable clinical benefit, NCB) according to the described provided in this study (Pender et al.,

2021), was also enriched in patients with higher value of AID mutations, p=5.3e^{-10}, Figure 4B. The study from Hugo et al. (Hugo et al., 2016), also confirmed this favorable association between higher fraction of AID mutations (cut-off according to the median) in patients with better OS treated with ICI in the univariate model, with log-rank p=0.018, and also in the multivariate Cox model, HR=0.37 [95% CI 0.37-0.99], p=0.048, Figures 5A-B. Finally, the Braun et al. study (Braun et al., 2020), also allowed to validate this prognostic association in univariate analysis, p=0.011, and in a Cox model after adjusting by age, gender and the presence of PRBM1 mutations, HR=0.68 [95% CI 0.51-0.92] p=0.012, Figures 6A-B.

[161] Overall, all the studies confirmed the independent prognostic value of high fraction of AID mutations according to the median in the univariate and multivariate analyses.

CLAIMS

1. An *in vitro* method for identifying a subject with cancer as being susceptible to respond to a treatment with an immune checkpoint inhibitor (ICI) or for prognosing survival of a subject with cancer and being treated with ICI, the method comprising assessing the fraction of AID-related mutations in a sample, wherein the fraction of AID-related mutations is the ratio of the number of AID-related mutations over the total number of mutated single nucleotide variants.

2. The *in vitro* method according to claim **1**, wherein an AID-related mutation is a mutation falling into an AID hotspot sequence, wherein said AID hotspot sequence includes the nucleic sequence WRCY or its reverse RGYW.

3. The *in vitro* method according to claim **2**, wherein the AID hotspot sequence includes AACC, AACT, AGCC, AGCT, TACC, TACT, TGCC, TGCT or the reverse TTGG, TTGA, TCGG, TCGA, ATGG, ATGA, ACGG, ACGA.

4. The in *vitro* method according to any one of claims **1** to **3**, wherein the sample is a tumor tissue.

5. The *in vitro* method according to any one of claims **1** to **4**, wherein said method is for identifying a subject with cancer as being susceptible to respond to a treatment with an ICI.

6. The *in vitro* method according to any one of claims **1** to **4**, wherein said method is for prognosing survival of a subject with cancer and being treated with ICI.

7. The *in vitro* method according to any one of claims **1** to **6**, wherein the ICI is selected from the group comprising an inhibitor of PD-1, an inhibitor of PD-L1, an inhibitor of CTLA-4 and a combination thereof.

8. The *in vitro* method according to any one of claims **1** to **7**, wherein the cancer is selected from the group comprising melanoma, non-small-cell lung carcinoma (NSCLC), renal cell carcinoma, head and neck cancers, merkel-cell carcinoma,

gastric cancer, small-cell lung carcinoma (SCLC), Hodgkin lymphoma, breast cancer, cervical cancer, colorectal cancer, endometrial cancer, hepatocellular cancer, esophageal cancer, mesothelioma, MSI (microsatellite instability)-high solid tumors, TMB (tumor mutation burden)-high tumors, breast cancer and urothelial carcinoma.

9. The *in vitro* method according to any one of claims **1** to **8**, the method further comprising the step of comparing the fraction of AID-related mutations with a reference value.

10. The *in vitro* method according to claim **9**, wherein the reference value is the median of the fractions of AID-related mutations measured in a reference population.

11. The *in vitro* method according to claim **9**, wherein the reference value is a decile of the fractions of AID-related mutations measured in a reference population.

12. The *in vitro* method according to claim **10** or claim **11**, wherein the reference population is a population of subjects having or having had a cancer, which are or have been treated with ICI, and which respond or have responded to ICI.

13. The *in vitro* method according to any one of claims **9** to **12**, wherein a fraction of AID-related mutation above the reference value is indicative of a subject as being susceptible to respond to a treatment with ICI, or prognosed with a high survival.

ABSTRACT

ACTIVATION-INDUCED CYTIDINE DEAMINASE AS A NEW BIOMARKER

The present invention relates to an activation-induced cytidine deaminase (AID) as a new biomarker for cancer.

Figure of abstract: Fig. 1

1/17

Cox−P Multivariate analysis (n = 249)
HR
95% CI
P value
Variables
Mutations
Binary (TMB Top 20% of each histology)
1.428
0.621−3.285
1.07 ×10−6
Binary (AICDA mutations Top 50% of all histologies)
0.201
0.088−0.458
2.29 ×10−2
Age Group
<30 (reference)
31−50
0.272
0.068−1.092
3.66 ×10−1
50−60
0.708
0.181−2.766
9.99 ×10−1
61−70
0.62
0.168−2.291
5.91 ×10−1
>71
0.338
0.084−1.358
4.53 ×10−1
Sex
Female (reference)
Male
0.846
0.485−1.478
2.4 ×10−2
0.062
0.125
0.250
0.500
1.00
2.00
<−−Better OS−−−−−Hazard Ratio−−−−− Worse OS−−>

FIG. 1

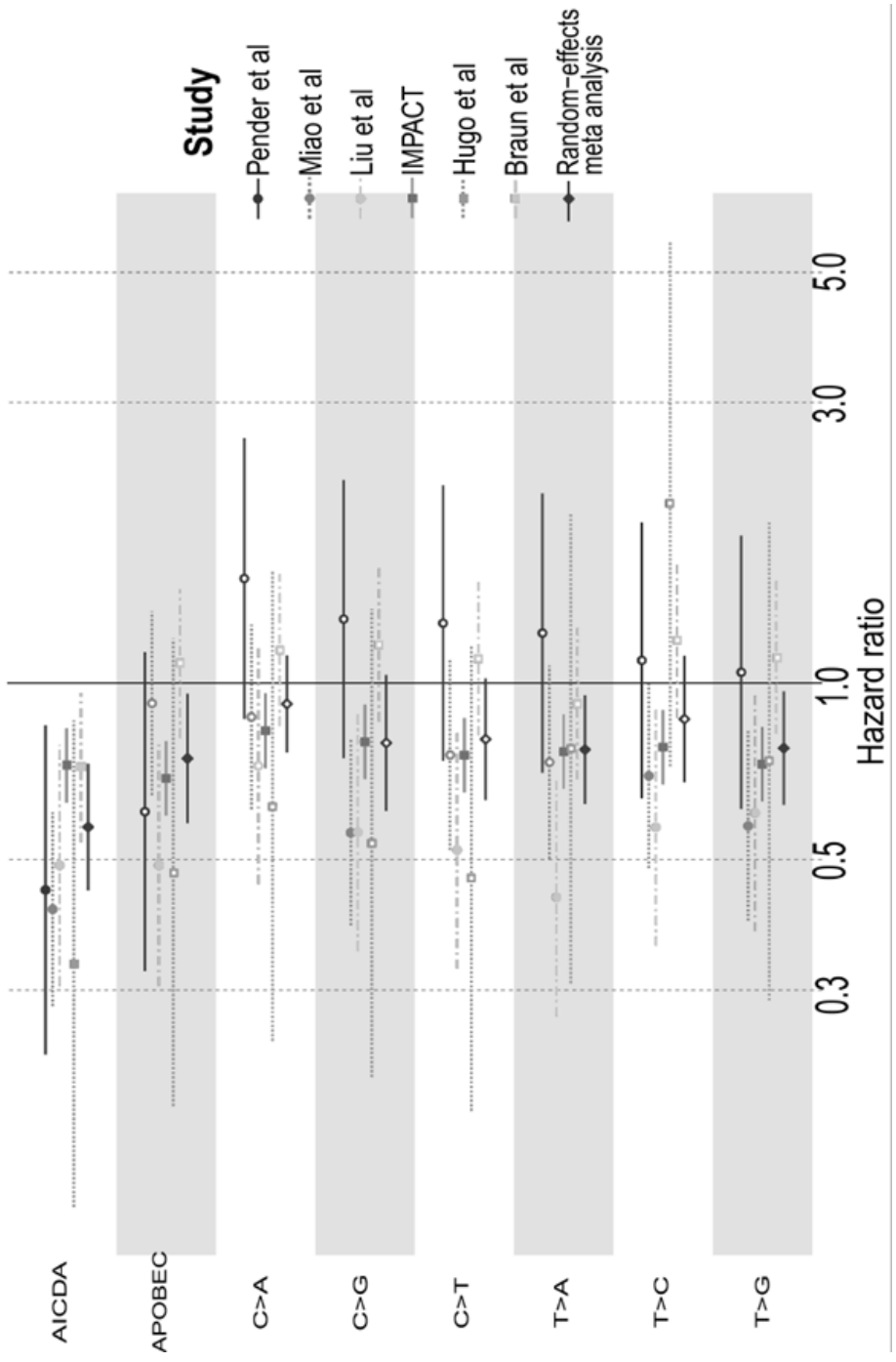

Study
Pender et al
Miao et al
Liu et al
IMPACT
Hugo et al
Braun et al
Random−effects meta analysis
AICDA
APOBEC
C>A
C>G
C>T
T>A
T>C
T>G
0.3
0.5
1.0
3.0
5.0
Hazard ratio

FIG. 2A

Cox–P Multivariate analysis (n = 1472)
HR
95% CI
P value
Variables
Mutations
Binary (TMB Top 20% of each histology)
0.795
0.661–0.956
1.5×10-2
Binary (AICDA mutations Top 50% of all histologies)
0.738
0.627–0.867
2.31×10-4
Binary (APOBEC mutations Top 50% of all histologies)
0.849
0.725–0.993
4.12×10-2
Age Group
<30 (reference)
31–50
1.174
0.729–1.891
4.07×10-1
50–60
1.226
0.757–1.985
9.24×10-1
61–70
0.977
0.606–1.575
6.74×10-1
>71
1.107
0.69–1.775
5.1×10-1
Sex
Female (reference)
Male
0.85
0.732–0.987
3.3×10-2
0.50
0.75
1.0
1.5
2.0
<---Better OS------Hazard Ratio------ Worse OS-->

FIG. 2B

Cox−P Multivariate analysis (n = 1472)

Variables	HR	95% CI	No. of Patients	Cutoff	P value
Cancer Type/Mutations					
All samples in cohort/TMB	0.681	0.574−0.808	1472		1.03×10^{-5}
All samples in cohort/AICDA	0.665	0.572−0.774	1472	0	1.25×10^{-7}
All samples in cohort/APOBEC	0.734	0.634−0.85	1472	0	3.76×10^{-5}
Melanoma/TMB	0.705	0.451−1.102	287	35.002	1.25×10^{-1}
Melanoma/AICDA	0.511	0.351−0.744	287	0	4.55×10^{-4}
Melanoma/APOBEC	0.621	0.427−0.903	287	0	1.27×10^{-2}
Cancer of Unknown Primary/TMB	0.372	0.143−0.969	76	15.799	4.28×10^{-2}
Cancer of Unknown Primary/AICDA	0.403	0.174−0.934	76	0	3.42×10^{-2}
Cancer of Unknown Primary/APOBEC	0.411	0.19−0.89	76	0	2.4×10^{-2}

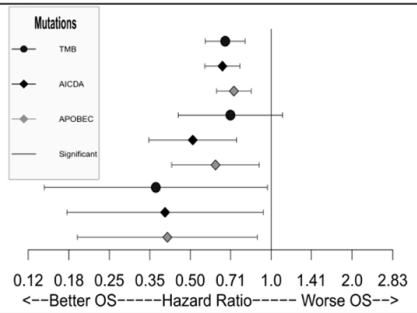


FIG. 2C

5/17

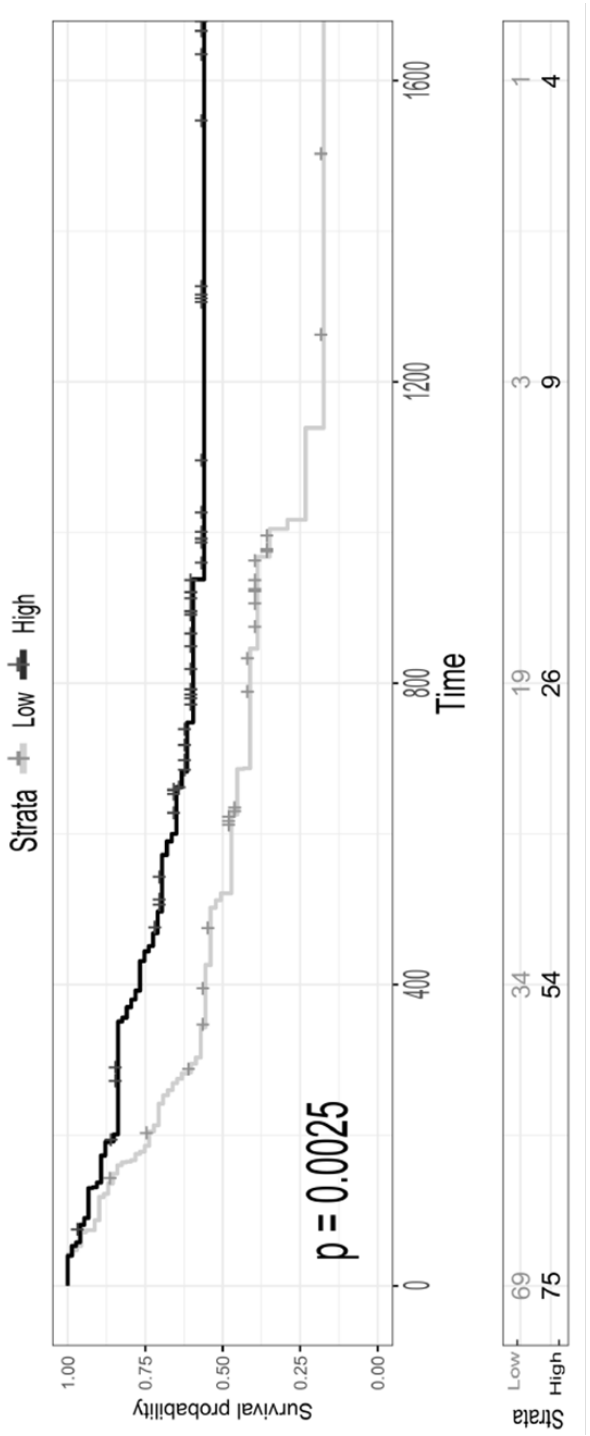

Strata
Low
High
p = 0.0025
Survival probability
1.00
0.75
0.50
0.25
0.00
0
400
800
1200
1600
Time
Strata
Low
High
69
75
34
54
19
26
3
9
1
4

FIG. 3A

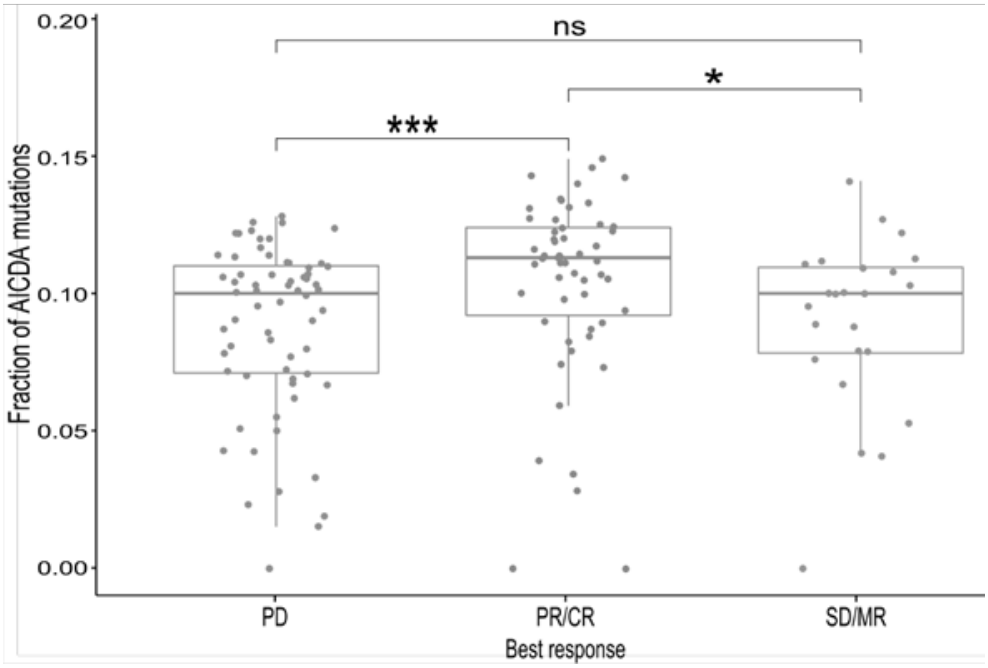

ns
*
***
0.20
0.15
0.10
0.05
0.00
Fraction of AICDA mutations
PD
PR/CR
SD/MR
Best response

FIG. 3B

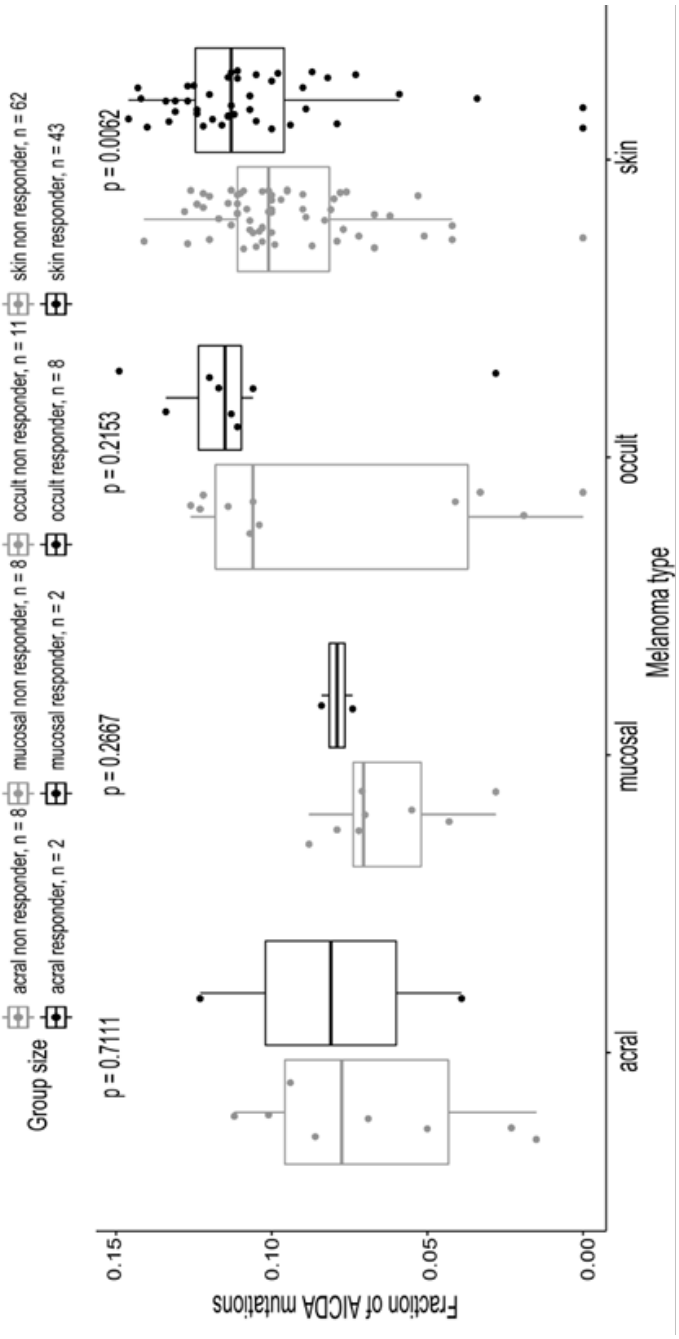

Group size
acral non responder, n = 8
acral responder, n = 2
mucosal non responder, n = 8
mucosal responder, n = 2
occult non responder, n = 11
occult responder, n = 8
skin non responder, n = 62
skin responder, n = 43
p = 0.7111
p = 0.2667
p = 0.2153
p = 0.0062
Fraction of AICDA mutations
0.15
0.10
0.05
0.00
acral
mucosal
occult
skin
Melanoma type

FIG. 3C

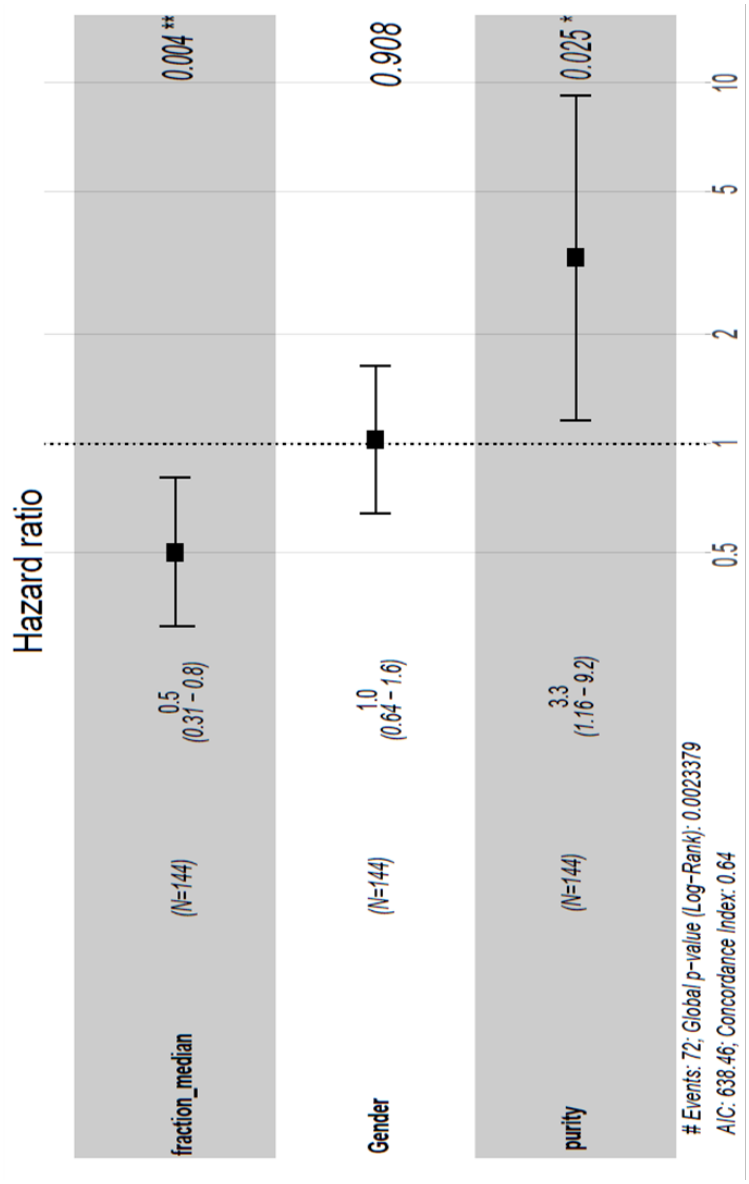

Hazard ratio
fraction_median
(N=144)
0.5
(0.31 – 0.8)
0.004 **
Gender
(N=144)
1.0
(0.64 – 1.6)
0.908
purity
(N=144)
3.3
(1.16 – 9.2)
0.025 *
0.5
1
2
5
10
Events: 72; Global p-value (Log-Rank): 0.0023379
AIC: 638.46; Concordance Index: 0.64

FIG. 3D

9/17

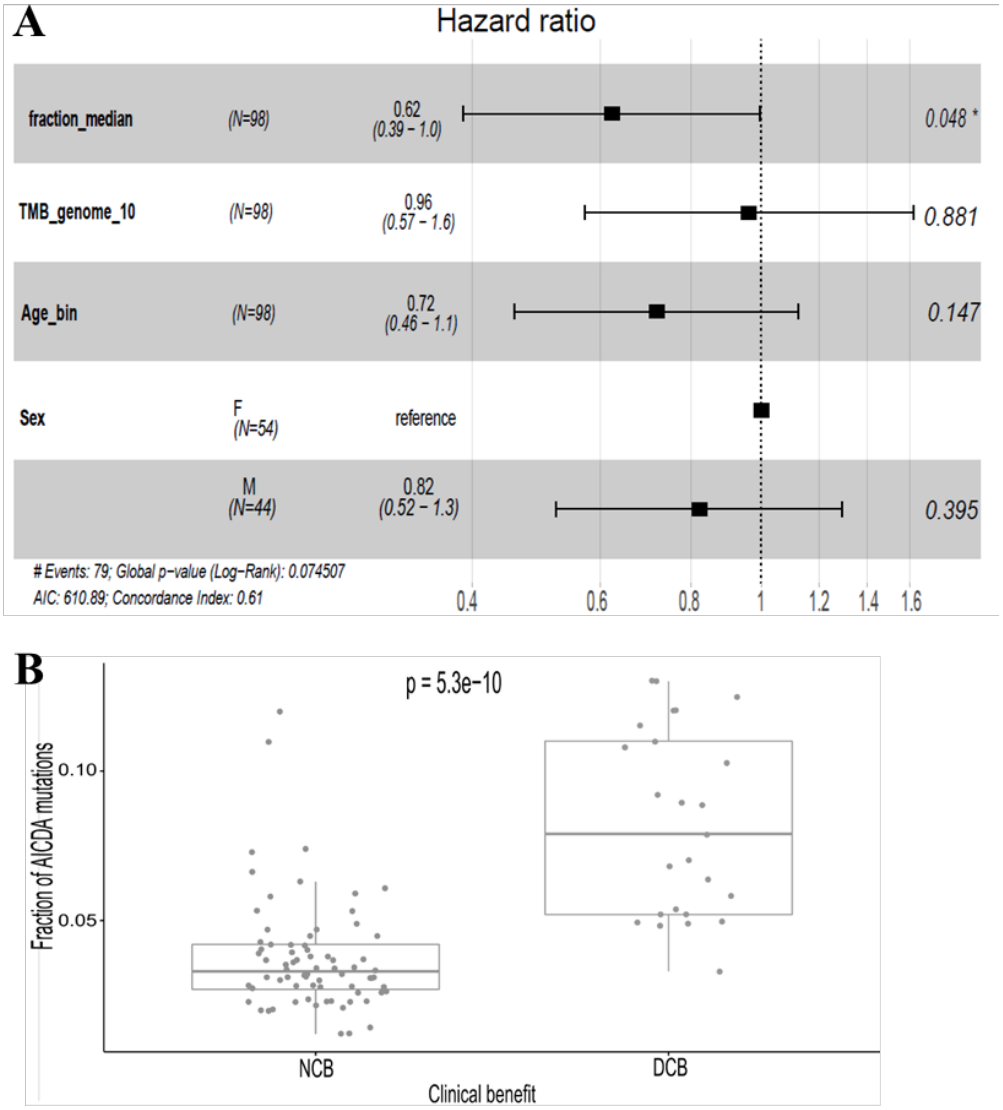

A
Hazard ratio
fraction_median
(N=98)
0.62
(0.39 – 1.0)
0.048 *
TMB_genome_10
(N=98)
0.96
(0.57 – 1.6)
0.881
Age_bin
(N=98)
0.72
(0.46 – 1.1)
0.147
Sex
F
(N=54)
reference
M
(N=44)
0.82
(0.52 – 1.3)
0.395
Events: 79; Global p−value (Log−Rank): 0.074507
AIC: 610.89; Concordance Index: 0.61
0.4
0.6
0.8
1
1.2
1.4
1.6
B
p = 5.3e−10
Fraction of AICDA mutations
0.10
0.05
NCB
DCB
Clinical benefit

FIG. 4

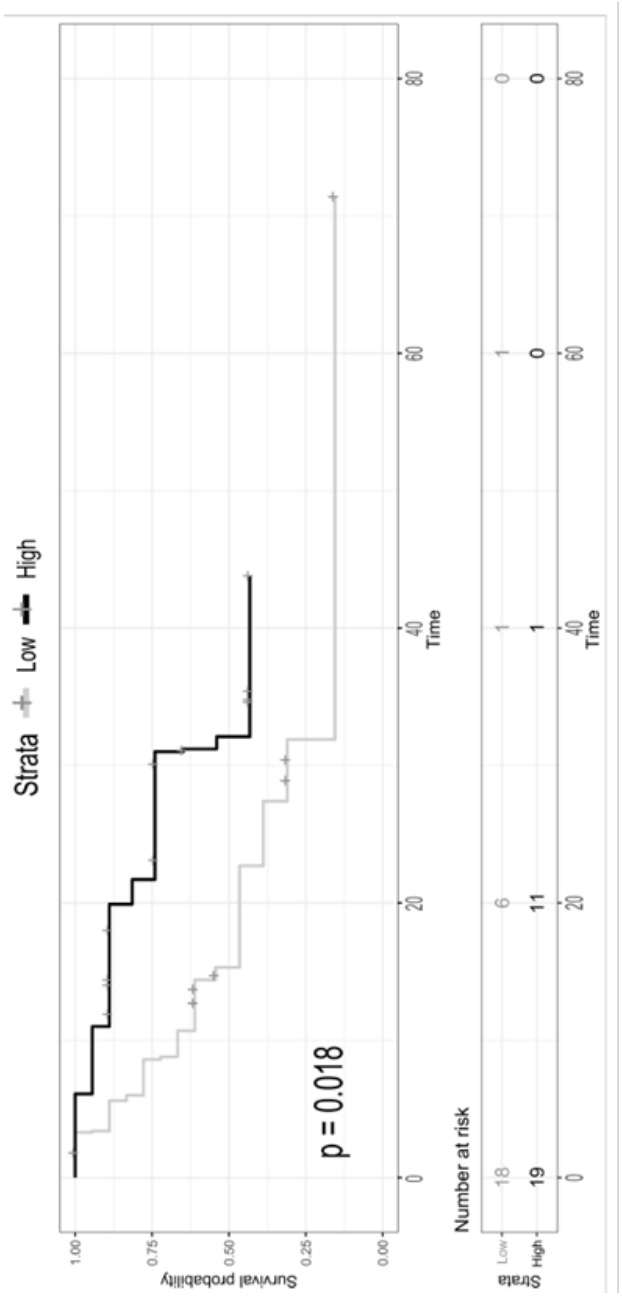

Strata
Low
High
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.018
0
20
40
60
80
Time
Number at risk
Strata
Low
High
18
19
6
11
1
0
1
0
0
0
0
20
40
60
80
Time

FIG. 5A

11/17

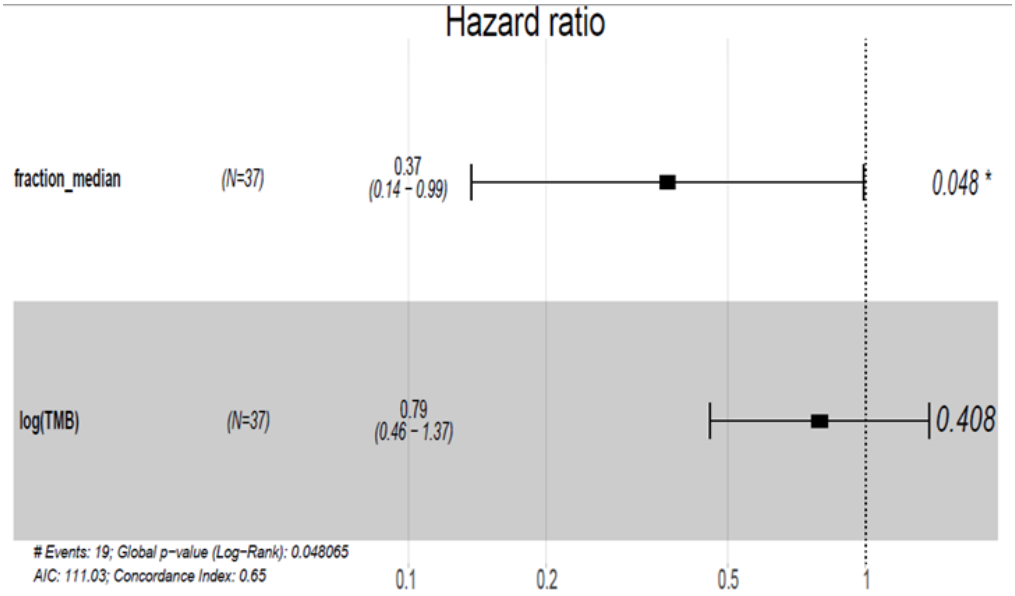

Hazard ratio
fraction_median
(N=37)
0.37
(0.14 - 0.99)
0.048 *
log(TMB)
(N=37)
0.79
(0.46 - 1.37)
0.408
Events: 19; Global p−value (Log−Rank): 0.048065
AIC: 111.03; Concordance Index: 0.65
0.1
0.2
0.5
1

FIG. 5B

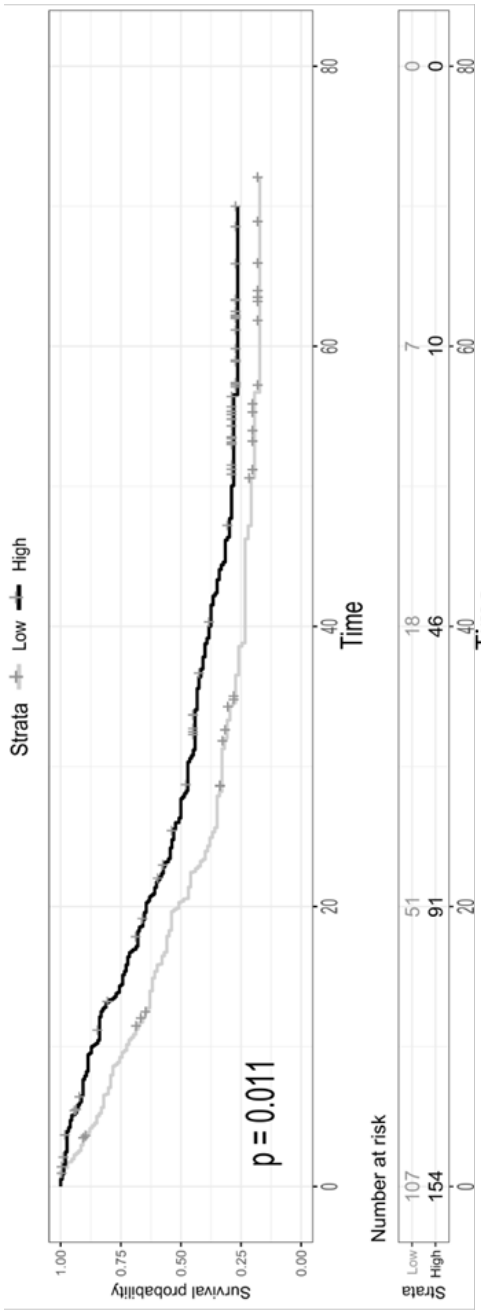

Strata
Low
High
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.011
0
20
40
60
80
Time
Number at risk
Low
High
107
154
51
91
18
46
7
10
0
0

FIG. 6A

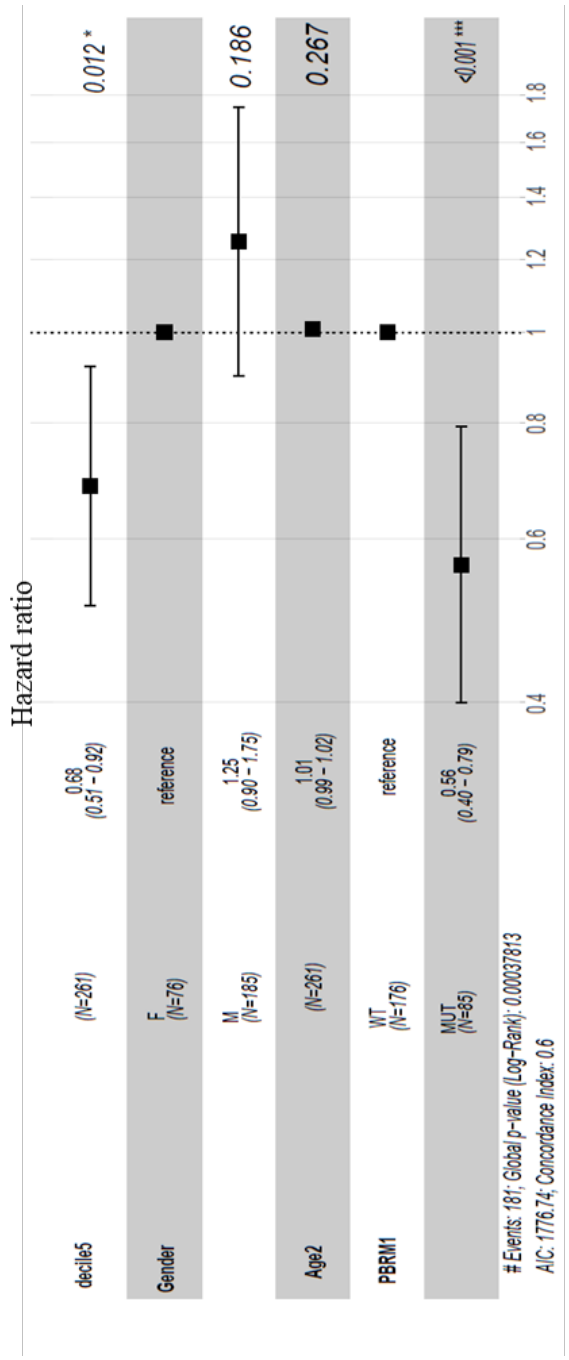

Hazard ratio
decile5
(N=261)
0.68
(0.51 – 0.92)
0.012 *
Gender
F
(N=76)
reference
M
(N=185)
1.25
(0.90 – 1.75)
0.186
Age2
(N=261)
1.01
(0.99 – 1.02)
0.267
PBRM1
WT
(N=176)
reference
MUT
(N=85)
0.56
(0.40 – 0.79)
<0.001 ***
Events: 181; Global p–value (Log–Rank): 0.00037813
AIC: 1776.74; Concordance Index: 0.6
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8

FIG. 6B

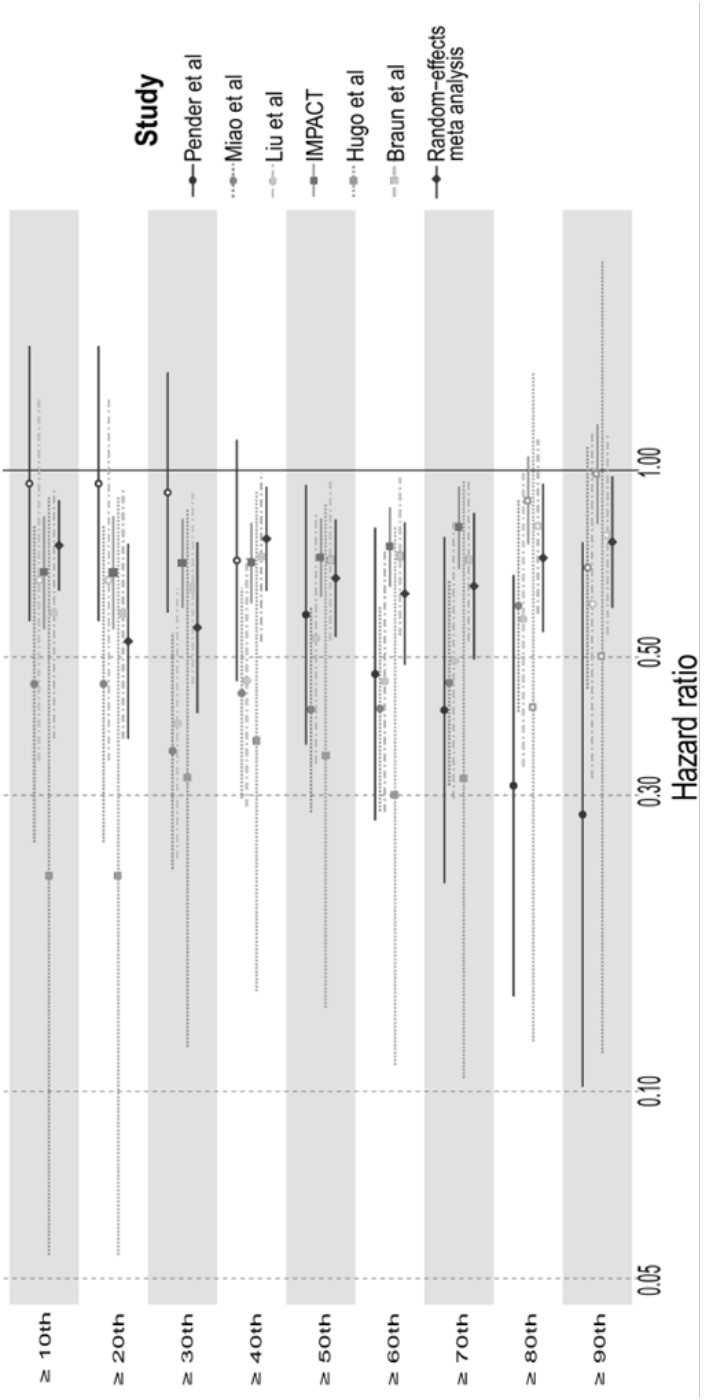

Study
Pender et al
Miao et al
Liu et al
IMPACT
Hugo et al
Braun et al
Random−effects meta analysis
≥ 10th
≥ 20th
≥ 30th
≥ 40th
≥ 50th
≥ 60th
≥ 70th
≥ 80th
≥ 90th
0.05
0.10
0.30
0.50
1.00
Hazard ratio

FIG. 7

15/17

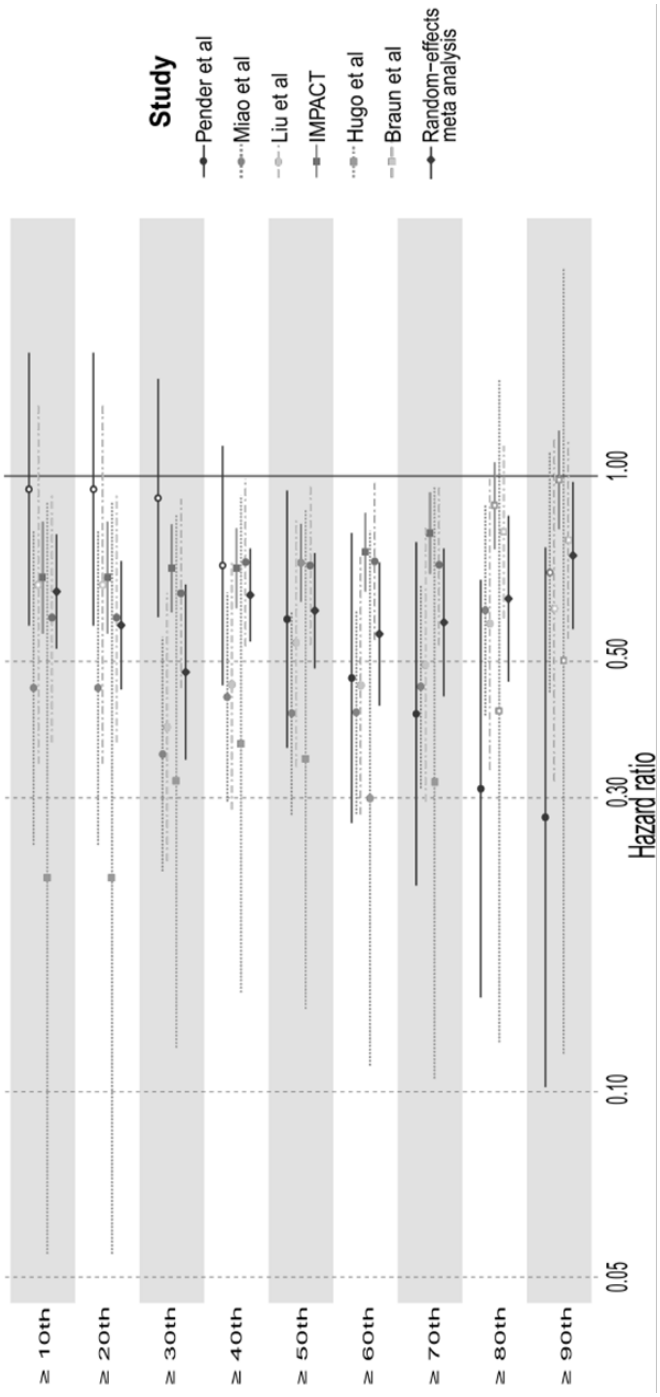

Study
Pender et al
Miao et al
Liu et al
IMPACT
Hugo et al
Braun et al
Random−effects meta analysis
≥ 10th
≥ 20th
≥ 30th
≥ 40th
≥ 50th
≥ 60th
≥ 70th
≥ 80th
≥ 90th
0.05
0.10
0.30
0.50
1.00
Hazard ratio

FIG. 8

Cox–P Multivariate analysis (n = 1472)
HR
95% CI
P value
Variables
Mutations
Binary (TMB >= 10)
0.633
0.516–0.776
1.07×10⁻⁵
Binary (AICDA mutations Top 50% of all histologies)
0.821
0.693–0.973
2.29×10⁻²
Binary (APOBEC mutations Top 50% of all histologies)
0.932
0.792–1.097
3.99×10⁻¹
Age Group
<30 (reference)
31–50
1.2
0.745–1.933
3.66×10⁻¹
50–60
1.249
0.771–2.022
9.99×10⁻¹
61–70
1
0.62–1.611
5.91×10⁻¹
>71
1.138
0.71–1.825
4.53×10⁻¹
Sex
Female (reference)
Male
0.841
0.724–0.978
2.4×10⁻²
0.50
0.75
1.0
1.5
2.0
<---Better OS------Hazard Ratio------ Worse OS-->

FIG. 9A

17/17

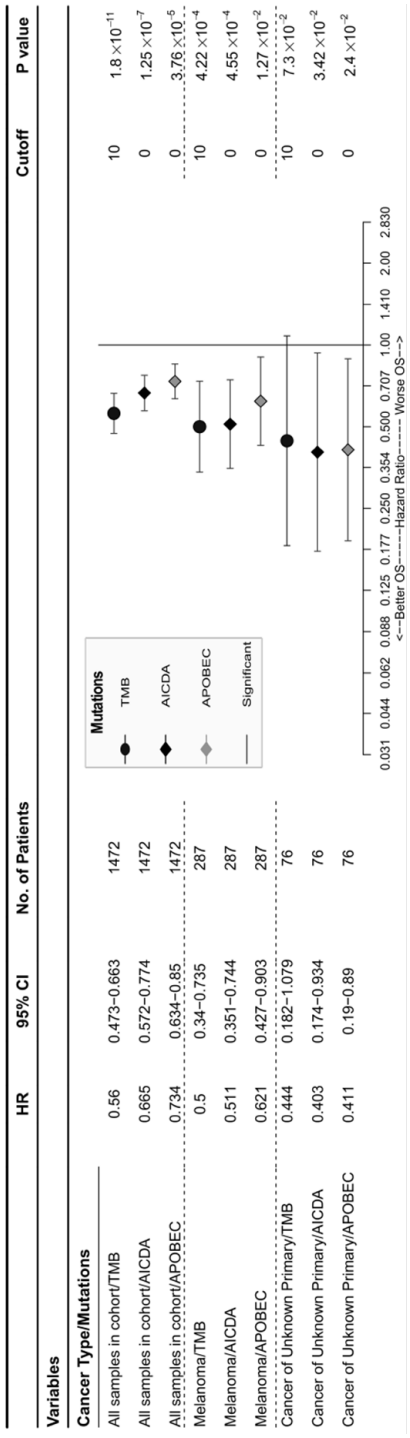

Cox–P Multivariate analysis (n = 1472)
HR
95% CI
No. of Patients
Cutoff
P value
Variables
Cancer Type/Mutations
All samples in cohort/TMB 0.56 0.473–0.663 1472 10 1.8 ×10-11
All samples in cohort/AICDA 0.665 0.572–0.774 1472 0 1.25 ×10-7
All samples in cohort/APOBEC 0.734 0.634–0.85 1472 0 3.76 ×10-5
Melanoma/TMB 0.5 0.34–0.735 287 10 4.22 ×10-4
Melanoma/AICDA 0.511 0.351–0.744 287 0 4.55 ×10-4
Melanoma/APOBEC 0.621 0.427–0.903 287 0 1.27 ×10-2
Cancer of Unknown Primary/TMB 0.444 0.182–1.079 76 10 7.3 ×10-2
Cancer of Unknown Primary/AICDA 0.403 0.174–0.934 76 0 3.42 ×10-2
Cancer of Unknown Primary/APOBEC 0.411 0.19–0.89 76 0 2.4 ×10-2
Mutations
TMB
AICDA
APOBEC
Significant
0.031 0.044 0.062 0.088 0.125 0.177 0.250 0.354 0.500 0.707 1.00 1.410 2.00 2.830
<---Better OS------Hazard Ratio------ Worse OS-->

FIG. 9B

Taylor & Francis
Taylor & Francis Group

Expert Review of Precision Medicine and Drug Development

Personalized medicine in drug development and clinical practice

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tepm20

Treating central nervous system lymphoma in the era of precision medicine

Ytel Garcilazo-Reyes , Maria-José Ibáñez-Juliá , Isaias Hernández-Verdin , Ludovic Nguyen-Them , Nadia Younan , Caroline Houillier , Khê Hoang-Xuan & Agusti Alentorn

To cite this article: Ytel Garcilazo-Reyes , Maria-José Ibáñez-Juliá , Isaias Hernández-Verdin , Ludovic Nguyen-Them , Nadia Younan , Caroline Houillier , Khê Hoang-Xuan & Agusti Alentorn (2020): Treating central nervous system lymphoma in the era of precision medicine, Expert Review of Precision Medicine and Drug Development, DOI: 10.1080/23808993.2020.1777853

To link to this article: https://doi.org/10.1080/23808993.2020.1777853

Accepted author version posted online: 02 Jun 2020.
Published online: 16 Jun 2020.

EXPERT REVIEW OF PRECISION MEDICINE AND DRUG DEVELOPMENT
https://doi.org/10.1080/23808993.2020.1777853

Taylor & Francis
Taylor & Francis Group

REVIEW

Treating central nervous system lymphoma in the era of precision medicine

Ytel Garcilazo-Reyes[a], Maria-José Ibáñez-Juliá[a,b], Isaias Hernández-Verdin[c], Ludovic Nguyen-Them[a,b], Nadia Younan[a,c], Caroline Houillier[a,d], Khê Hoang-Xuan[a,c,d] and Agusti Alentorn[a,c]

[a]APHP, Department of Neurology-2, Groupe Hospitalier Pitié Salpêtrière, Paris, France; [b]Department of Neurology, CH Perpignan, Perpignan, France; [c]Sorbonne Université, Paris, France; [d]Réseau Expert National LOC (Lymphomes Oculo-Cérébraux), Groupe Hospitalier Pitié Salpêtrière, Paris, France

ABSTRACT

Introduction: Primary central nervous system lymphoma (PCNSL) is a rare extra-nodal non-Hodgkin lymphoma that in the vast majority of cases belongs to diffuse large B-cell lymphoma (DLBCL) histology. The standard first-line treatment is based on high-dose methotrexate (HD-MTX) regimens. However, the majority of patients will relapse, leading to a poor prognosis of the disease.

Areas covered: Reviewed are the potential new therapeutic approaches in PCNSL. With the advent of tailored treatment, immunomodulators and immunotherapies are appearing as new promising therapeutic approaches for this orphan disease. This review seeks to summarize the novel approaches currently under evaluation.

Expert opinion: The therapeutic management of PCNSL is rapidly evolving with the description of PCNSL molecular alterations. However, due to the rarity of this disease, phase III clinical trials using new therapeutic drugs are still lacking. In addition, the vast majority of newly diagnosed PCNSL affect elderly patients, and specific and adapted clinical trials for this fragile population are warranted.

Currently, the use of targeted therapies or immune-mediated treatments is only studied in relapsed/refractory (R/R) PCNSL, but the use of these approaches as a first-line treatment (compared with HD-MTX) could also be used as new promising approaches to decrease the toxicity associated with MTX regimens.

ARTICLE HISTORY
Received 12 December 2019
Accepted 1 June 2020

KEYWORDS
Bruton's tyrosine kinase inhibitors; immune checkpoint inhibitors; immune-related therapies; primary central nervous system lymphoma; lenalidomide; mTOR inhibitors; targeted therapies

1. Introduction

PCNSL is a rare subtype of lymphoma, representing roughly 2% of primary central nervous system (CNS) tumors [1]. It is a challenging form of non-Hodgkin lymphoma restricted to the CNS or eyes [2]. Close to 90% of PCNSL cases are Diffuse Large B-cell Lymphomas (DLBCLs); the remainder are T-cell lymphomas, indolent B-cell lymphomas, Burkitt's lymphomas, and poorly characterized low-grade lymphomas. Despite the important therapeutic improvement in the past decades, the prognosis of PCNSL remains poor with a median overall survival (OS) of 26 months [3]. However, a substantial number of patients may hope to be cured since the 5-year and 10-year survival rates for PCNSL are 29.9% and 22.2%, respectively [4]. About one third of patients are refractory to first-line treatment, and up to 60% of the patients will eventually relapse [5]. There is no standard chemotherapy regimen for R/R PCNSL. The prognosis is poor even with salvage therapy, the median progression-free survival (PFS) after recurrence is around 2.2 months (range, 0–29.6) and the median OS is approximately 3.5 months (range, 0–29.6) [6].

DLBCLs are a heterogenous group of tumors with different pathogenic mechanisms [7]. Therefore, to understand the efficacy and failures of the new drugs is important to dissect the molecular background of this disease. Recently, the molecular alterations that characterize the PCNSL have been deeply described but using small numbers of patients.

1.1. Lymphomagenesis and tumor microenvironment

DLBCLs of the CNS are recognized as mature post-germinal B cells, this is the called activated B-cell (ABC) immunophenotype. This phenotype traduces cell positivity to CD19, CD20, CD79a-, BCL-6, and MUM1/IRF4. These cells carry rearranged and somatically mutated immunoglobulin (Ig) genes with evidence of ongoing somatic hypermutation (SHM) [8]. For a brief recall, when B lymphocytes are activated, they undergo rapid proliferation and simultaneously initiate two genome remodeling reactions, termed SHM and class-switch recombination (CSR). SHM introduces point mutations in the variable region of Ig genes, which can increase antibody affinity, whereas CSR is a DNA deletion event that replaces one Ig constant region gene for another [9].

PCNSLs exhibit several types of low-frequency genetic alterations including somatic mutations, copy number alterations (CNAs), and chromosomal rearrangements [10]. These genetic alterations may lead to activation of the B-cell receptor (BCR), Toll-like receptor (TLR), and NF-kβ pathways in more than 90% of cases [11].

CONTACT Agusti Alentorn agusti.alentorn@aphp.fr AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière - Charles Foix, Paris 75013, France

Article highlights

- Half of PCNSL present hotspot somatic mutations in *MYD88* and *CD79B* leading to activation of TCR and BCR signaling pathways.
- Ibrutinib, an inhibitor of the BCR pathway, is a promising therapeutic agent in PCNSL.
- The use of immune-related regimens is an interesting approach but its efficacy in PCNSL should be confirmed in future clinical trials.

Aberrant SHM is not confined to its physiological targets but extends to other genes that have been implicated in tumorigenesis [8,12]. Somatic mutations particularly lead to oncogenic TLR and BCR signaling pathway alterations. Mutations affect the Ig genes in 38% of cases and *BCL6* in 17–47% of cases, *MYD88* in 38–75% of cases, *BCL2* in 42–90% of the cases and *CD79B* in approximately 20% of cases [3,13–15]. Oncogenic mutations of the TLR adaptor protein MYD88 and/or *NFKBIZ* copy gain are common genetic features of Epstein Bar Virus (EBV) PCNSL [11]. *MYD88* mutations have also been identified in vitreous aspirates in primary vitreoretinal lymphomas (PVRL) [13]. The crosstalk between TLR and BCR signaling pathway allows an interesting target inhibition of pathway components such as IRAK1/4, IRF4, and/or BTK. A subset of PCNSLs exhibits activating *CARD11* mutations in association with *MYD88* and *CD79B*, which may limit the efficacy of proximal BCR pathway inhibitors [16].

CNAs may lead to genomic instability and less favorable outcomes. PCNSL can alter the p53 pathway via upstream of *CDKN2A* loss [17]. It has been described the presence of 9p24.1/PD-L2 copy gains in association with 9p21.3/*CDKN2A* copy loss and increased genomic instability [11]. CNAs of 9p24.1 and translocations of 9p24.1 lead to overexpression of the programmed cell death 1 receptor (PD-1) ligands, PD-L1, and PD-L2 [4]. Besides immune evasion, another mechanism affecting the prognosis is a suppressed tumor immune microenvironment.

Recent studies have highlighted the significant roles of the tumor microenvironment in driving tumor progression and the development of chemo/radio-resistance. Lymphoma cells in an important percentage of the cases (>60%) seem to lack cell surface expression of the MHC class I complex, which is necessary for the recognition by cytotoxic T lymphocytes [10,18]. Tumor-associated macrophages/microglia (TAMs) are a major stromal cell component in PCNSL [19], these TAMs preferentially express PD-L1 and that high expression of PD-L1 by TAMs shows a trend toward a correlation with a poor outcome [20,21].

1.2. Treatment strategies

1.2.1. Immuno-chemotherapy

HD-MTX (>3 g/m^2) has been combined with other chemotherapeutic agents such as cytarabine, procarbazine, temozolomide, ifosfamide, etoposide, and thiotepa. According to the latest PCNSL guidelines, HD-MTX in combination with cytarabine is recommended for a combination treatment for newly diagnosed PCNSL [22].

The role of the chimeric monoclonal anti-CD20 antibody, rituximab, in association with the standard therapy is still controversial, and conclusions of two prospective studies show different results [23–25]. The international IELSG32 trial had three induction chemotherapy arms. The first arm HD-MTX and cytarabine, the second arm had the same regimen combined with rituximab, and the third arm had the same as the second in addition to thiotepa, known as MATRix regimen. It was demonstrated that the addition of rituximab alone and the MATRix regimen increased CR rates compared to the first arm. The addition of rituximab also proved better objective response rates (ORR). The addition of thiotepa to rituximab further improved ORR rate and outcome [25]. These encouraging results were not confirmed by the group of Bromberg, their patients received induction therapy either with HD-MTX therapy, associated with carmustine, etoposide, and rituximab or induction chemotherapy without rituximab. No improvement in the event-free survival (EFS), OS and response rate (RR) in the arm of rituximab was demonstrated, apparently patients 60 years and younger may be the subgroup that could benefit [23]. Therefore, a further analysis focusing on the benefit by age group for the use of rituximab may result interesting.

1.2.2. Target therapies (BCR/TLR pathway and PI3 K/mTOR pathway)

Knowing that BCR signaling plays a major role in these tumors. It can be targeted upstream or downstream. Upstream inhibition could target the spleen tyrosine kinase (SYK), phosphatidylinositol – 4,5 – bisphosphate 3 – kinase (PI3 K), Bruton Tyrosine Kinase (BTK), or Interleukin-1 receptor-associated kinase (IRAK). Downstream the pathway could be inhibited by immunomodulatory drugs (IMIDs) like thalidomide and its analogs lenalidomide and pomalidomide which inhibit IRF4, or inhibitors of mucosa-associated lymphoid tissue lymphoma translocation protein 1 (MALT1) [26].

Ibrutinib, a small molecule that binds permanently to BTK, resulted in a promising candidate drug for assessment in PCNSL [27]. BTK links BCR activity to NF-κβ and is essential for the survival of ABC lines with chronic active BCR signaling [16]. The efficacy of ibrutinib is noteworthy with high response rates and interesting PFS. A preclinical study demonstrated the high level of ibrutinib brain distribution, which supports the clinical potential value of this drug in the PCNSL treatment armamentarium [28]. Ibrutinib demonstrated significant clinical activity particularly in tumors harboring both *CD79B* and *MYD88* hotspot mutations. This was demonstrated in a phase Ib study, which objective was to assess the response rate according to the different molecular subtypes of DLBCL, PFS, and OS, and the association of ibrutinib response with genomic aberrations that alter BCR and NF-κβ signaling in ABC DLBCL (*CD79B, MYD88, CARD11*, and *TNFAIP3*). Overall responses (OR) were observed in 25% of the patients, median PFS and OS were 1.64 months and 6.41 months, respectively [29].

Several studies have also assessed the tolerance and the potential efficacy of ibrutinib in PCNSL either alone or in combination. In monotherapy, response was observed in 77% of 13 patients with PCNSL with a median PFS of

4.6 months. This was an open-label dose-escalation study of ibrutinib in patients with PCNSL or secondary central nervous system lymphoma (SCNSL), with R/R disease. The maximum tolerated dose (MTD) of ibrutinib was 840 mg demonstrating higher concentrations in cerebrospinal fluid (CSF) on day 29. Ibrutinib also confirmed higher response rates in PCNSL than reported for DLBCL outside the CNS, suggesting a divergent molecular pathogenesis. Mutations of *CARD11* were found in the only patient with resistance and in some patients with partial response (PR). Patients with complete response (CR) did not have *CD79B* mutations either [30]. A phase Ib conducted in 18 PCNSL patients, ibrutinib was used as a first-line treatment (alone or in combination) and/or as a treatment of R/R disease. Ninety-four percent showed tumor reductions with ibrutinib alone, including patients having PCNSL with *CD79B* and/or *MYD88* mutations, and 86% of evaluable patients achieved CR with DA-TEDDi-R (temozolomide, etoposide, doxil, dexamethasone, ibrutinib, and rituximab) with a median PFS of 15.3 months [31]. An important limitation of this article was the fact that it was not possible to know if some patients might have eventually reached CR on ibrutinib monotherapy alone.

Another recent phase Ib study combining ibrutinib with high-dose methotrexate and rituximab regimen in R/R PCNSL reported responses even in PCNSL without *MYD88* or *CD79B* mutations. In this study, six out of nine (67%) patients achieved CR, and these patients mostly corresponded to an ABC subtype (85.7%), they reached a median PFS of 9.2 months [32]. The LOC network has conducted a phase II trial enrolling 52 patients with refractory and relapsed PCNSL (including 8 PVRL) using 560 mg daily dosed ibrutinib. The primary endpoint was the disease control (DC) rate, including complete and unconfirmed complete response (CR and uCR), partial response (PR) and stable disease (SD) after two cycles of treatment. Ibrutinib was detected in the CSF, DC was achieved in 70% of patients, treatment failed in 13 patients, ORR was observed in 59%, with a median PFS of 4.8 months and OS of 19 months [33].

A shared complication seen in the studies of ibrutinib was the risk of aspergillosis (pulmonary and/or CNS). It is worth mentioning a higher frequency in combination regimens compared to monotherapy with ibrutinib. The hypothesis behind these infections is that this treatment impairs fungal immune surveillance, a deficit that may be exacerbated by co-administration of dexamethasone and/or chemotherapy. In their murine model (Btk knockout and wild-type mice) aspergillosis was linked to BTK-dependent fungal immunity. Macrophages provide the first line of defense against fungi and the exposure of macrophage TLRs to fungal pathogens initiates downstream signaling, including activation of BTK, promoting adaptive immune responses. Macrophage TLR activation is required for immunity and inflammatory responses to *Aspergillus fumigatus* [34].

On the other hand, it has been suggested the interest in blocking the PI3 K/mammalian target of rapamycin (PI3 K/mTOR) pathway to overcome resistance in *CD79B* mutated tumors. The PI3 K signaling pathway plays a critical role in oncogene-mediated tumor growth and proliferation and has regulatory functions in cell survival, apoptosis, protein synthesis, and glucose metabolism [35,36]. Down in this signalization pathway, it is found a serine-threonine protein kinase, mTOR. Temsirolimus is an mTOR inhibitor, that has been studied in R/R PCNSL, mTOR is now recognized as a unique and important target for cancer therapeutics. There is only one prospective study that investigated the mTOR inhibition in R/R PCNSL using temsirolimus. This phase II nonrandomized, open-label study used temsirolimus as a single-agent with a two-stage design. In the first stage, patients were treated with temsirolimus 25 mg intravenously once per week, if no common toxicity criteria grades 3 to 4 were observed, all following patients were treated with 75 mg once per week. Korfel and colleagues demonstrated a high radiographic response of 54%, but a low median PFS of only 2.1 months, suggesting a transient effect [37]. Even less encouraging results were reported with buparlisib. Buparlisib is an oral pan-PI3 K inhibitor that had shown antitumor activity in lymphoma cell lines and induced apoptosis in DLBCL [38]. However, in a phase II trial, there was a response rate of 25% a median PFS of 39 days an 100% of relapses [39].

1.2.3. Immunotherapy

Nivolumab, pembrolizumab (anti-PD1), and durvalumab (anti-PDL1), the so-called immune checkpoint inhibitors (ICIs), have also been assessed [40]. Today no prospective trial has been completed, and only case reports or series of cases have been described. Terziev et al. reported the first documented case of treatment with nivolumab after high-dose chemotherapy with autologous stem cell transplant (HD-CT/ASCT) in a PCNSL patient, this patient showed a sustained CR [41]. Nivolumab, a human IgG4 monoclonal antibody that targets PD-1 and blocks engagement of the PD-1 ligands, was used in four patients with refractory or recurrent PCNSL and in one SCNSL and a response was seen in all of them [42]. Results of prospective trials are expected in the near future [NCT02857426]. The use of nivolumab in combination with dendritic cell vaccination has been described [43]. Dendritic cell vaccination is a cancer immunotherapy in which dendritic cells are cultured and loaded with tumor antigen *ex vivo* activate T-cell to attack tumor cells by presenting tumor antigen [44].

Pembrolizumab also binds to the PD-1 receptor, blocking both immune-suppressing ligands, PD-L1 and PD-L2, from interacting with PD-1 to help restore T-cell response and immune response [45]. Two pembrolizumab clinical trials [NCT02779101, NCT03012620] are currently evaluating ORR in patients treated with pembrolizumab for relapsed PCNSL after MTX-based first-line therapy. There is evidence of expression of PD-L1 and/or PD-L2 in a subset of non-Hodgkin lymphomas as well as in the tumor microenvironment, making this pathway a promising target [46]. A phase II clinical trial [NCT03212807] using durvalumab plus lenalidomide for R/R DLBCL, PCNSL, and PTL is currently ongoing. Durvalumab is a high-affinity human IgG1 monoclonal antibody that binds to PD-1 and CD80, blocking PD-L1, but not PD-L2. The primary endpoint of this study is to evaluate the ORR after 6 months of follow up. Because the PD-L1 ligand is located on the tumor

cells, a PD-L1 inhibitor should penetrate both the blood-brain barrier (BBB) and the blood-tumor barrier effectively. Albeit challenging, due to the low prevalence of these lymphomas and the uncertain pharmacokinetic properties of these antibodies with regard to the BBB, their clinical use should be actively be explored in this population [46].

Adoptive cell therapy (ACT) has been tested in different types of R/R B-cell cancers [47–51]. This immunotherapy strategy includes the use of chimeric antigen receptor (CAR)-T cells. CAR-T cells treatment is based on the incorporation of T cells that have been genetically engineered to express a CAR for the pan–B-cell CD-19 antigen [52]. Single-center studies of anti-CD19 CAR-T cells in refractory DLBCL have shown encouraging results, with rates of complete remission of more than 50% and some durable remissions in a subset of patients [48,53]. One explanation for the different response rates among tumor types is that CAR-T functionality may be inhibited by an immunosuppressive tumor microenvironment. All available CAR-T cells trials have excluded patients with CNS involvement. Neelapu et al. and Schuster et al. reported with axicabtagene ciloleucel and tisagenlecleucel, respectively, important ORR of 82% and 56% in patients followed by systemic DBLCL. It is worth to mention the neurologic adverse reactions informed in the previous studies (12–28%) using this strategy and sensitize the neurologist and neuro-oncologist in charge. CAR-T cell therapy has demonstrated the ability to cross the BBB and induce responses in the CNS [53,54]. A case report of primary refractory DLBCL involving the brain parenchyma achieved CR with JCAR017, a CD19-directed CAR-T cell product [51]. There is currently an ongoing phase II clinical multi-cohort in adults with aggressive B-cell non-Hodgkin lymphoma, including R/R PCNSL [NCT03484702]. Recently ongoing clinical trials mentioned above are summarized in Table 1.

1.2.4. Immunomodulators

The group of IMIDS is headed by lenalidomide and pomalidomide, oral agents derived from thalidomide, with antiproliferative properties. Lenalidomide has antiproliferative properties that modify the microenvironment and activates cytotoxic T cells and NK cells [55]. Lenalidomide has been studied as monotherapy, in

Table 1. Ongoing clinical trials in PCNSL.

Drug(s)	NCT identifier	Type of study	Population	Primary endpoint	Secondary endpoint(s)
Ibrutinib Ibrutinib + HD-MTX	NCT02315326	Phase I/II	R/R PCNSL and SCNSL	Establish MTD of ibrutinib * MTD of Ibrunib + HD-MTX	Safety/tolerability PFS DOR
Ibrutinib/copanlisib	NCT03581942	Phase Ib/ II	R/R PCNSL	MTD and ORR	Adverse effects PFS DOR OS
Ibrutinib/rituximab/ lenalidomide	NCT03703167	Phase Ib	R/R PCNSL and SCNSL	MTD	PFS
Nivolumab	NCT02857426	Phase II	Relapsed/Refractory PCNSL or relapsed refractory PTL	ORR	PFS OS DOR
Pembrolizumab	NCT02779101	Phase II	Relapsed PCNSL after MTX-based first line therapy	ORR	-
Pembrolizumab	NCT03012620	Phase II	Multi-cohort study with a dedicated cohort of R/R PCNSL	ORR	PFS OS DOR
Nivolumab/ pomalidomide	NCT03798314	Phase I	r/r PCNSL and PVRL	MTD	ORR PFS
Rituximab+Lenalidomide + Nivolumab	NCT03558750	Phase I/II	R/R Non-Germinal Center Type DBCL or PCNSL	MTD and toxicity Efficacy	Tolerability, Time to progression, ORR with and without MYD88
Durvalumab + Lenalidomide	NCT03212807	Phase II	R/R EBV+ associated DLBCL Subtypes, PCNSL and PTL.	ORR	PFS OS
Pomalidome/ dexametasone	NCT01722305	Phase I	Dose-escalation study for R/R PCNSL	Stablish MTD Efficacy and safety	DOR Safety
Lenalidomide/rituximab	NCT01956695	Phase II	R/R PCNSL	Efficacy measured by ORR	Safety Duration of response PFS OSS Quality of life
Buparlisib (BKM120)	NCT02301364	Phase II	R/R PCNSL and R/R SCNSL	PFS	AE OS ORR
PQR309	NCT02669511	Phase II	R/R PCNSL	ORR	AE
Temsirolimus	NCT00942747	Phase II	R/R PCNSL	ORR	Safety Time to progression Penetration in CSF
CAR-T cells‡	NCT03484702	Phase II	R/R PCNSL and other aggressive B-NHL	ORR	Safety PFS OS

PCNSL: Primary Central Nervous System Lymphoma, PTL: Primary Testicular Lymphoma, SCNSL: Secondary Nervous System Lymphoma, DBCL: Diffuse B-cell Lymphoma, R/R: Relapsed/Refractory, EBV: Epstein Barr Virus, B-NHL: B-cell non-Hodgkin lymphoma; MTX: Methotrexate, HD-MTX: High-Dose Methotrexate, AE: adverse events, PFS: progression-free survival, ORR: overall response rate, OS: overall survival, DOR: duration of response, MTD: Maximum-tolerated dose, CSF: cerebrospinal fluid, ‡Autologous T-cells expressing anti-CD19 chimeric antigen receptor.

combination with rituximab and as maintenance therapy. In a proof of concept that with aging T cells, B cells, dendritic cells, and NK cells change, this drug has been assessed in elderly patients, proving that it suppresses apoptosis of stimulated T cells via interleukin 2-dependent mechanisms [56] and that it has a moderate activity with a good tolerability in this population.

As a single agent, lenalidomide has been used in patients with recurrent PCNSL, with a description of two patients achieving CR [57]. In a phase I study by Rubenstein et al., it was reported an MTD of lenalidomide of 15 mg/day in a 21-day/28 schema, and radiographic responses were observed in 64% (9/14 R/R PCNSL) with a median PFS of 6 months [58]. A phase II French study recently suggested the efficacy of lenalidomide in 50 R/R PCNSL patients (lenalidomide 20 mg/21 days was combined with rituximab), an ORR at the end of induction of 35.6% was observed, responding patients followed a maintenance with lenalidomide, and the median PFS was 7.8 months (3.9–11.3), the OS was 17.7 months (12.9 to not reached), and a good toxicity profile was observed as well [59]. Low dose lenalidomide (5–10 mg/day) as maintenance treatment in patients older than 70 years, demonstrated to be well tolerated, and with a median overall follow up of 31.6 months the median PFS had not been reached [60].

Pomalidomide is a third-generation IMID that has shown a good CNS penetration [61]. In a phase I study, Tun et al. assessed the efficacy of pomalidomide in R/R PCNSL and in PVRL patients. The ORR reported was 48% and a median PFS of 9 months for responders, and it was well tolerated in terms of side effects [62].

1.2.5. Associating strategies

As aforementioned, we discussed the association of IMIDS and monoclonal antibodies, BTKi, and conventional chemotherapy, but a question that may arise is the possibility of associating immunotherapy strategies, for example, CD19 CAR-T cells, IMIDS, and/or ICIs. Indoleamine 2,3-dioxygenase (IDO) is an intracellular enzyme induced by inflammatory mediators that intervenes on the metabolism of immunosuppressive metabolites, which blocks antigen-specific T-cell proliferation and induces T-cell death. IDO´s activity antagonizes CD19 CAR-T cell therapy [63]. One study proved the overexpression of PD-L1 and IDO1 by macrophage/microglia in PCNSL tissues. Suggesting that the expression of immunosuppressive molecules, including PD-1 ligands and IDO1, by macrophage/microglia may be involved in immune evasion of lymphoma cells [64]. Rubenstein provides evidence for the activation of the IDO pathway in CNS lymphomas, raising the possibility that IDO may contribute to early resistance to lenalidomide [58]. Further evaluation of this treatment approach is needed.

2. Conclusion

There has been an enormous progress in the last decades in understanding this pathology. HD-MTX-based chemotherapies remain the standard induction strategy. There are still controversies in the induction therapies, notably regarding the use of rituximab. Forconsolidation therapy, novel therapies are encouraged. Understanding PCNSL biology has allowed to incorporate target therapies and immunotherapies, but with every new drug, challenges will appear. In a global vision, these strategies have not reached durable desired responses and have short PFS. However, among these molecules, ibrutinib and lenalidomide have demonstrated promising results. It is important to highlight that ibrutinib could be associated with unexpected adverse effects like a potential arrhythmogenic risk and a higher risk of aspergillosis. IMIDs might also help to improve outcomes, particularly in the elderly. The role of ICIs is not yet clear. Some small case series of R/R PCNSL had shown a high rate of radiological responses. Corticosteroids increase the risk of life-threatening infections and may interfere with the efficacy of ICIs. CAR-T cell therapy has been proved to penetrate the CNS, the concern of neurotoxicity exists, there are no completed clinical trials for PCNSL, so far. Current ongoing trials will give more resources to the clinician in the years coming.

3. Expert Opinion

The standard of care for newly diagnosed PCNSL patients is based on HD-MTX regimens. However, there is a broad range of unanswered therapeutic questions that should be considered. For instance, whether it is necessary to consolidate HD-MTX to improve the response rate or whether the efficacy of immune-related or targeted therapies will be different if they were used as first-line treatments instead of at disease recurrence. Several clinical trials are currently ongoing to try to answer this clinically relevant question.

It is worth mentioning that the therapeutic management of PCNSL is rapidly evolving with the description of PCNSL molecular alterations. However, due to the rarity of this disease, phase III clinical trials using new therapeutic drugs are still lacking. In addition, the vast majority of the newly diagnosed PCNSL affect elderly patients and specific and adapted clinical trials for this fragile population should be conducted.

Currently, the use of targeted therapies or immune-mediated treatments is mainly studied in R/R PCNSL, but the use of these approaches as a first-line treatment (compared with HD-MTX) should be encouraged as the new promising approaches may decrease the toxicity associated with HD-MTX regimens.

Another important point with potential therapeutic consequences is the role of the brain tumor microenvironment in the evolution and therapeutic response of PCNSL. In this line, the boost of immune response using ICIs, immunomodulators, or CAR-T cell therapy could be a promising therapeutic approach.

Furthermore, it is necessary to perform high-throughput molecular studies using a larger number of samples. The vast majority of molecular studies in PCNSL have been performed in small cohorts ($n < 25$) or using only a high-throughput approach (i.e. either whole-exome or RNA-seq) and large multi-omics studies are still lacking. Indeed, there is virtually no molecular alteration that has robustly been associated with the clinical evolution or the prognosis of PCNSL. Moreover, the comprehensive molecular portrait of PCNSL using multi-omics approaches could allow to identify new clinically relevant therapeutic or theranostic targets.

In the next years, there will certainly appear different clinical trials combining targeted therapies associated with immune-

related therapeutic agents in PCNSL tailored according to the genetic and molecular background of this CNS lymphoma.

Funding

This research was funded in part by ANR-18-RHUS-0012 and PRT-K 16149 (DGOS and INCa).

Declaration of interest

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Reviewers disclosure

A reviewer on this manuscript has disclosed having an investigator initiator trial grant from Janssen on Ibrutinib as maintenance treatment in elderly patients with PCNSL. Peer reviewers on this manuscript have no other relevant financial relationships or otherwise to disclose.

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Ostrom QT, Gittleman H, Fulop J, et al. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2008–2012. Neuro Oncol. 2015;17(Suppl 4):iv1–iv62.
2. Batchelor TT, Grossman SA, Mikkelsen T, et al. Rituximab monotherapy for patients with recurrent primary CNS lymphoma. Neurology. 2011;76(10):929–930.
3. Mendez JS, Quinn OT, Kruchko C, et al. Changes in survival of primary central nervous system lymphoma based on a review of national databases over 40 years. JCO. 2017;35:2040.
4. Braggio E, Van Wier S, Ojha J, et al. Genome-wide analysis uncovers novel recurrent alterations in primary central nervous system lymphomas. Clin Cancer Res. 2015;21:3986–3994.
5. Fischer L, Jahnke K, Martus P, et al. The diagnostic value of cerebrospinal fluid pleocytosis and protein in the detection of lymphomatous meningitis in primary central nervous system lymphomas. Haematologica. 2006;91:429–430.
6. Langner-Lemercier S, Houillier C, Soussain C, et al. Primary CNS lymphoma at first relapse/progression: characteristics, management, and outcome of 256 patients from the French LOC network. Neuro Oncol. 2016;18:1297–1303.
7. Monti S, Savage KJ, Kutok JL, et al. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood. 2005;105:1851–1861.
8. Louis D, Ohgaki H, Wiestler O WHO Classification of Tumours of the Central Nervous System. 2016.
9. Chaudhuri J, Alt FW. Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. Nat Rev Immunol. 2004;4:541–552.
10. Basso K, Dalla-Favera R. Germinal centres and B cell lymphomagenesis. Nat Rev Immunol. 2015;15:172–184.
11. Chapuy B, Roemer MGM, Stewart C, et al. Targetable genetic features of primary testicular and primary central nervous system lymphomas. Blood. 2016;127:869–881.
•• **characterization of genetic signatures in PCNSL and PTL for possible target therapies.**
12. Maul RW, Gearhart PJ. AID and somatic hypermutation. Adv Immunol. 2010;105:159–191.
13. Kraan W, Horlings HM, van Keimpema M, et al. High prevalence of oncogenic MYD88 and CD79B mutations in diffuse large B-cell lymphomas presenting at immune-privileged sites. Blood Cancer J. 2013;3:e139.
14. Knittel G, Liedgens P, Korovkina D, et al. B-cell-specific conditional expression of Myd88p.L252P leads to the development of diffuse large B-cell lymphoma in mice. Blood. 2016;127:2732–2741.
15. Gonzalez-Aguilar A, Idbaih A, Boisselier B, et al. Recurrent mutations of MYD88 and TBL1XR1 in primary central nervous system lymphomas. Clin Cancer Res. 2012;18:5203–5211.
16. Davis RE, Ngo VN, Lenz G, et al. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. Nature. 2010;463:88–92.
17. Monti S, Chapuy B, Takeyama K, et al. Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B-cell lymphoma. Cancer Cell. 2012;22:359–372.
18. Challa-Malladi M, Lieu YK, Califano O, et al. Combined genetic inactivation of β2-microglobulin and CD58 reveals frequent escape from immune recognition in diffuse large B cell lymphoma. Cancer Cell. 2011;20:728–740.
19. Komohara Y, Jinushi M, Takeya M. Clinical significance of macrophage heterogeneity in human malignant tumors. Cancer Sci. 2014;105:1–8.
20. Hayano A, Komohara Y, Takashima Y, et al. Programmed cell death ligand 1 expression in primary central nervous system lymphomas: a clinicopathological study. Anticancer Res. 2017;37:5655–5666.
• **Analisis of PD-L1 expression in PCNSLs.**
21. Sasayama T, Tanaka K, Mizowaki T, et al. Tumor-associated macrophages associate with cerebrospinal fluid interleukin-10 and survival in Primary Central Nervous System Lymphoma (PCNSL). Brain Pathol. 2016;26:479–487.
22. Hoang-Xuan K, Bessell E, Bromberg J, et al. Diagnosis and treatment of primary CNS lymphoma in immunocompetent patients: guidelines from the European Association for Neuro-Oncology. Lancet Oncol. 2015;16(7):e322–332.
•• **Current european guidelines for the approach and treatment of PCNSL.**
23. Bromberg JEC, Issa S, Bakunina K, et al. Rituximab in patients with primary CNS lymphoma (HOVON 105/ALLG NHL 24): a randomised, open-label, phase 3 intergroup study. Lancet Oncol. 2019;20:216–228.
•• **Phase III study that proved no benefit of the addition of rituximab in the induction chemotherapy.**
24. Ferreri AJM, Cwynarski K, Pulczynski E, et al. Whole-brain radiotherapy or autologous stem-cell transplantation as consolidation strategies after high-dose methotrexate-based chemoimmunotherapy in patients with primary CNS lymphoma: results of the second randomisation of the International Extranodal Lymphoma Study Group-32 phase 2 trial. Lancet Haematol. 2017;4:e510–e523.
25. Ferreri AJM, Cwynarski K, Pulczynski E, et al. Chemoimmunotherapy with methotrexate, cytarabine, thiotepa, and rituximab (MATRix regimen) in patients with primary CNS lymphoma: results of the first randomisation of the International Extranodal Lymphoma Study Group-32 (IELSG32) phase 2 trial. Lancet Haematol. 2016;3:e217–227.
•• **Largest randomized trial comparing different induction chemo-immunotherapies in PCNSL, best outcomes in combination with rituximab.**
26. Grommes C, Nayak L, Tun HW, et al. Introduction of novel agents in the treatment of primary CNS lymphoma. Neuro Oncol. 2019;21:306–313.
27. Löw S, Batchelor TT. Primary central nervous system lymphoma. Semin Neurol. 2018;38:86–94.
28. Goldwirt L, Beccaria K, Ple A, et al. Ibrutinib brain distribution: a preclinical study. Cancer Chemother Pharmacol. 2018;81:783–789.
29. Wilson WH, Young RM, Schmitz R, et al. Targeting B cell receptor signaling with ibrutinib in diffuse large B cell lymphoma. Nat Med. 2015;21:922–926.
30. Grommes C, Pastore A, Palaskas N, et al. Ibrutinib unmasks critical role of bruton tyrosine kinase in primary CNS lymphoma. Cancer Discov. 2017;7:1018–1029.
• **PCNSL trial using ibrutinib in R/R PCNSL.**
31. Lionakis MS, Dunleavy K, Roschewski M, et al. Inhibition of B cell receptor signaling by ibrutinib in primary CNS lymphoma. Cancer Cell. 2017;31:833–843.e5.

• **Ibrutinib as monotherapy and in combination in PCNSL and assess of the role of BTK in *A. fumigatus* immune surveillance.**
32. Grommes C, Tang SS, Wolfe J, et al. Phase 1b trial of an ibrutinib-based combination therapy in recurrent/refractory CNS lymphoma. Blood. 2019;133:436–445.
33. Soussain C, Choquet S, Blonski M, et al. Ibrutinib monotherapy for relapse or refractory primary CNS lymphoma and primary vitreoretinal lymphoma: final analysis of the phase II "proof-of-concept" iLOC study by the lymphoma study association (LYSA) and the French oculo-cerebral lymphoma (LOC) network. Eur J Cancer. 2019;117:121–130.
• **Phase II trial confirming the activity of Ibrutinib as single agent in PCNSL.**
34. Herbst S, Shah A, Mazon Moya M, et al. Phagocytosis-dependent activation of a TLR9-BTK-calcineurin-NFAT pathway co-ordinates innate immunity to Aspergillus fumigatus. EMBO Mol Med. 2015;7:240–258.
35. Fingar DC, Blenis J. Target of rapamycin (TOR): an integrator of nutrient and growth factor signals and coordinator of cell growth and cell cycle progression. Oncogene. 2004;23:3151–3171.
36. Katso R, Okkenhaug K, Ahmadi K, et al. Cellular function of phosphoinositide 3-kinases: implications for development, homeostasis, and cancer. Annu Rev Cell Dev Biol. 2001;17:615–675.
37. Korfel A, Schlegel U, Herrlinger U, et al. Phase II trial of temsirolimus for relapsed/refractory primary CNS lymphoma. J Clin Oncol. 2016;34:1757–1763.
• **Temsirolimus monotherapy demonstrated to be active in PCNSL patients failled to demonstrate sustained response.**
38. Zang C, Eucker J, Liu H, et al. Inhibition of pan-class I phosphatidyl-inositol-3-kinase by NVP-BKM120 effectively blocks proliferation and induces cell death in diffuse large B-cell lymphoma. Leuk Lymphoma. 2014;55:425–434.
39. Grommes C, Pentsova E, Nolan C, et al. Phase II study of single agent buparlisib in recurrent/refractory primary (PCNSL) and secondary CNS lymphoma (SCNSL). Ann Oncol. 2016;27:vi106.
40. Juárez-Salcedo LM, Sandoval-Sus J, Sokol L, et al. The role of anti-PD-1 and anti-PD-L1 agents in the treatment of diffuse large B-cell lymphoma: the future is now. Crit Rev Oncol Hematol. 2017;113:52–62.
41. Terziev D, Hutter B, Klink B, et al. Nivolumab maintenance after salvage autologous stem cell transplantation results in long-term remission in multiple relapsed primary CNS lymphoma. Eur J Haematol. 2018;101:115–118.
42. Nayak L, Iwamoto FM, LaCasce A, et al. PD-1 blockade with nivolumab in relapsed/refractory primary central nervous system and testicular lymphoma. Blood. 2017;129:3071–3073.
• **Nivolumab as immunotherapy active in R/R PCNSL patiens.**
43. Furuse M, Nonoguchi N, Omura N, et al. Immunotherapy of nivolumab with dendritic cell vaccination is effective against intractable recurrent primary central nervous system lymphoma: a case report. Neurol Med Chir. 2017;57:191–197.
44. Kalinski P, Muthuswamy R, Urban J. Dendritic cells in cancer immunotherapy: vaccines and combination immunotherapies. Expert Rev Vaccines. 2013;12:285–295.
45. Sharma P, Allison JP. The future of immune checkpoint therapy. Science. 2015;348:56–61.
46. Andorsky DJ, Yamada RE, Said J, et al. Programmed death ligand 1 is expressed by non-hodgkin lymphomas and inhibits the activity of tumor-associated T cells. Clin Cancer Res. 2011;17:4232–4244.
47. Turtle CJ, Hanafi L-A, Berger C, et al. Immunotherapy of non-hodgkin's lymphoma with a defined ratio of CD8+ and CD4+ CD19-specific chimeric antigen receptor-modified T cells. Sci Transl Med. 2016;8:355ra116.
48. Schuster SJ, Svoboda J, Chong EA, et al. Chimeric antigen receptor T cells in refractory B-cell lymphomas. N Engl J Med. 2017;377:2545–2554.
49. Schuster SJ, Bishop MR, Tam CS, et al. Tisagenlecleucel in adult relapsed or refractory diffuse large B-cell lymphoma. N Engl J Med. 2019;380:45–56.
50. Neelapu SS, Locke FL, Bartlett NL, et al. Axicabtagene ciloleucel CAR T-cell therapy in refractory large B-cell lymphoma. N Engl J Med. 2017;377:2531–2544.
51. Abramson JS, McGree B, Noyes S, et al. Anti-CD19 CAR T cells in CNS diffuse large-B-cell lymphoma. N Engl J Med. 2017;377:783–784.
52. Maus MV, Grupp SA, Porter DL, et al. Antibody-modified T cells: cARs take the front seat for hematologic malignancies. Blood. 2014;123:2625–2635.
53. Kochenderfer JN, Dudley ME, Kassim SH, et al. Chemotherapy-refractory diffuse large B-cell lymphoma and indolent B-cell malignancies can be effectively treated with autologous T cells expressing an anti-CD19 chimeric antigen receptor. J Clin Oncol. 2015;33:540–549.
54. Grupp SA, Kalos M, Barrett D, et al. Chimeric antigen receptor–modified T cells for acute lymphoid leukemia. N Engl J Med. 2013;368:1509–1518.
55. Zhu D, Corral LG, Fleming YW, et al. Immunomodulatory drugs revlimid (lenalidomide) and CC-4047 induce apoptosis of both hematological and solid tumor cells through NK cell activation. Cancer Immunol Immunother. 2008;57:1849–1859.
56. Huang M-C, Greig NH, Luo W, et al. Preferential enhancement of older human T cell cytokine generation, chemotaxis, proliferation and survival by lenalidomide. Clin Immunol. 2011;138:201–211.
57. Houillier C, Choquet S, Touitou V, et al. Lenalidomide monotherapy as salvage treatment for recurrent primary CNS lymphoma. Neurology. 2015;84:325–326.
58. Rubenstein JL, Geng H, Fraser EJ, et al. Phase 1 investigation of lenalidomide/rituximab plus outcomes of lenalidomide maintenance in relapsed CNS lymphoma. Blood Adv. 2018;2:1595–1607.
59. Ghesquieres H, Chevrier M, Laadhari M, et al. Lenalidomide in combination with intravenous rituximab (REVRI) in relapsed/refractory primary CNS lymphoma or primary intraocular lymphoma: a multicenter prospective 'proof of concept' phase II study of the French Oculo-Cerebral lymphoma (LOC) Network and the Lymphoma Study Association (LYSA)†. Annals of Oncology : Official Journal of the European Society for Medical Oncology. 2016;127(4):621–628.
• **Demonstration of the efficacy of rituximab plus lenalidomide in PCNSL patients.**
60. Vu K, Mannis G, Hwang J, et al. Low-dose lenalidomide maintenance after induction therapy in older patients with primary central nervous system lymphoma. Br J Haematol. 2019;186:180–183.
61. Li Z, Qiu Y, Personett D, et al. Pomalidomide shows significant therapeutic activity against CNS lymphoma with a major impact on the tumor microenvironment in murine models. PLoS ONE. 2013;8:e71754.
62. Tun HW, Johnston PB, DeAngelis LM, et al. Phase 1 study of pomalidomide and dexamethasone for relapsed/refractory primary CNS or vitreoretinal lymphoma. Blood. 2018;132:2240–2248.
63. Ninomiya S, Narala N, Huye L, et al. Tumor indoleamine 2,3-dioxygenase (IDO) inhibits CD19-CAR T cells and is downregulated by lymphodepleting drugs. Blood. 2015;125:3905–3916.
64. Miyasato Y, Takashima Y, Takeya H, et al. The expression of PD-1 ligands and IDO1 by macrophage/microglia in primary central nervous system lymphoma. J Clin Exp Hematop. 2018;58:95–101.

Increased brain delivery of anti-programmed death-ligand 1 using low-intensity pulsed ultrasound-mediated blood-brain barrier opening is associated with increased anti-tumor efficacy and microglia activation in glioblastoma mouse models

Mohammed Ahmed[1,2*], Isaias Hernández-Verdin[1], Nolwenn Lemaire[1], Emie Quissac[1], Coralie L Guerin[3], Lea Guyonnet[3], Noël Zahr[4], Laura Mouton[5], Mathieu Santin[5], Alexandra Petiet[5], Charlotte, Schmitt[6], Guillaume Bouchoux[6], Michael Canney[6], Marc Sanson[7], Maite Verreault[1], Alexandre Carpentier[6,7], Ahmed Idbaih[7*]

[1]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital de la Pitié Salpêtrière, Paris, France

[2]Faculté de Médecine Paris-Sud - Université Paris-Saclay, 91190, Saint-Aubin, France.

[3]Curie Institute, Cytometry Department, F-75006, Paris, France.

[4]Pharmacokinetics and Therapeutic Drug Monitoring unit, INSERM, CIC-1901, UMR ICAN 1166, Pitié-Salpêtrière Hospital, Sorbonne Université, AP-HP, F-75013 Paris, France.

[5]Centre de Neuroimagerie de Recherche, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital de la Pitié Salpêtrière, Paris, France

[6]CarThera, Institut du Cerveau et de la Moelle épinière (ICM), F-75013, Paris, France

[7]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, AP-HP, Hôpital de la Pitié Salpêtrière, DMU Neurosciences, Service de Neurologie 2-Mazarin, F-75013, Paris, France

*To whom correspondence should be addressed

Mohammed Ahmed, Pharm.D, Ph.D.

Institut du Cerveau - Paris Brain Institute-ICM,

Inserm, CNRS, APHP, Hôpital de la Pitié Salpêtrière

Paris, France

Tel: + 33 1 57 27 4485

Fax: +33 1 57 27 40 27

Email : mhia2@cam.ac.uk

Ahmed Idbaih, M.D, Ph.D.

Sorbonne Université,

Institut du Cerveau-Paris Brain Institute - ICM, Inserm,

CNRS, AP-HP, Hôpital de la Pitié Salpêtrière, DMU Neurosciences,

Service de Neurologie 2-Mazarin, F-75013,

Paris, France.

Tel: 01-42-16-03-85.

Fax: 01-42-16-04-18.

Email : ahmed.idbaih@aphp.fr

Abstract

Therapeutic antibodies targeting immune checkpoints have shown limited efficacy in clinical trials in glioblastoma (GBM) patients. Ultrasound-mediated blood-brain barrier (BBB) opening (UMBO) using low-intensity pulsed ultrasound (LIPU) improved drug delivery to the brain. We explored the safety and the efficacy of LIPU plus immune checkpoint inhibitors (ICIs) in preclinical models of GBM. BBB opening was performed using a 1 MHz preclinical ultrasound system in combination with 10µl/gram microbubbles. The penetration of programmed death-1 (anti-PD-1) and programmed death-ligand 1 (anti-PDL-1) checkpoint inhibitors were measured by nano-surface and molecular-orientation limited (nSMOL) proteolysis followed by liquid chromatography-mass spectrometry (LC-MS/MS) bioanalysis and immunohistochemistry.

The impact of repeated treatments on survival were determined. In syngeneic GBM-bearing immunocompetent mice, we showed that UMBO safely and repeatedly open the BBB. BBB opening was confirmed visually and microscopically using Evans's blue dye and magnetic resonance imaging. UMBO plus anti-PDL-1 was associated with a significant improvement of the overall survival compared to anti-PD-L1 alone. Using mass spectroscopy, we showed that the penetration of a therapeutic antibody can be increased by 28-fold when delivered intravenously compared to non-sonicated brains. Furthermore, we observed an enhancement of the of activated microglia percentage when combined with anti-PD-L1. Here, we report that the

combination of UMBO and anti-PD-L1 therapeutic antibody increases GBM-bearing mice's survival dramatically compared to their counterparts treated with anti-PD-L1 alone. Our study highlights the BBB as a limitation to overcome to increase the efficacy of ICIs in GBM and supports clinical trials combining UMBO and anti-PD-L1 in GBM patients.

Keywords: SonoCloud, GL261 mouse model, BBB, ultrasound-mediated drug delivery, glioblastoma, checkpoint inhibitor, PDL1

Importance of the Study

Here, we used orthotopic murine GBM models to investigate the therapeutic combination of PD-L1 checkpoint blockade with ultrasound-mediated blood-brain barrier opening (UMBO). Our data confirm the ability of UMBO to increase antibody delivery to the brain. Interestingly, this increased delivery was associated with an increased anti-tumor efficacy of anti-PD-L1. The antitumor effect of UMBO plus anti-PD-L1 could be explained in part by an increase in microglia activation. This work opens new avenues for the efficacy of UMBO plus ICIs in the treatment of GBM.

Introduction

Glioblastoma (GBM) is the most malignant primary brain tumor in adults, with a median overall survival of less than 18 months after initial diagnosis[1]. Despite significant efforts in the neuro-oncology field to develop new therapeutic alternatives, temozolomide (approved in 2005) remains today the standard first-line chemotherapy in GBM treatment[1,2]. For over five decades, research has been focused on developing new anti-cancer therapies for GBM, including anti-neoplastic agents[3], molecular targeted drugs[4], immunotherapeutic approaches[5], and angiogenesis inhibiting compounds[6]; however, the prognosis of patients has hardly improved and temozolomide remains the only chemotherapy shown to improve patient survival in randomized clinical trials[7].

The existence of the blood-brain barrier (BBB), which is specific to the blood vessels in the central nervous system (CNS), prevents most systemic therapeutic compounds from reaching the brain parenchyma and GBM cells[8] although it is disrupted in some areas (i.e. blood tumor barrier).

Several innovative strategies have been studied to enhance the delivery of chemotherapeutic agents and antibodies to the brain[8]. Ultrasound-mediated blood-brain barrier opening (UMBO) using low-intensity pulsed ultrasound (LIPU) has now been studied in preclinical[9] and clinical settings[10]. LIPU is delivered to the brain simultaneously with an intravenous injection

(IV) of micron-sized bubbles for a few minutes, allowing the microbubbles to oscillate. Microbubble oscillation produces a mechanical stretching on vessel walls that allows a transient BBB opening[11]. UMBO has shown a good safety profile for BBB opening in recurrent GBM patients[10,12] and is now being studied in dozens of clinical trials using a range of transcranial [13]or implantable ultrasound devices[14] for treating both primary and secondary brain tumors as well as neurodegenerative diseases[15]

The choice of therapeutic agents to deliver after UMBO is crucial and remains a point of discussion among researchers. Direct stimulation of the immune system with immune checkpoint inhibitors (ICI, *e.g.*, PD-1/PD-L1) showed promising effects alone or with other chemotherapies in multiple cancers. Ipilimumab was the first humanized anti-CTLA-4 approved to treat inoperable melanoma[16]. Five years later, Atezolizumab was the first humanized anti-PD-L1 approved to treat advanced or metastatic urothelial carcinoma[17]. PD-L1 proteins are expressed as surface molecules by cancerous cells such as GBM cells[18] and provide a tumor escape mechanism when bound to PD-1 proteins at the surface of activated T-lymphocytes leading to their exhaustion[19]. Despite their promise in other cancers, nivolumab (anti-PD-1) has shown no additional efficacy over bevacizumab in phase III clinical trials in recurrent GBM patients[20]. Similarly, Avelumab (anti-PD-L1) in combination with molecular targeted drugs did not improved outcome of GBM patients[21]

In the present study, we evaluated the effect of anti-PD-L1 and anti-CTLA-4 alone and in combination with UMBO in syngeneic Gl261 and Nfpp10 mouse models.

Materials and Methods

Cell culture and in *vivo* studies

GL261 cells were cultured in Dulbecco's modified essential medium (DMEM) supplemented with 10% fetal bovine serum and 1% Penicillin/Streptomycin. Cells were passaged twice weekly according to their confluence. Nfpp10-luciferase cell line (*NF1*, *PTEN*, and *TP53* deficient) as previously described[22] were maintained in culture using DMEM/F12 (Gibco; Life Technologies) culture medium supplemented with 1% penicillin-streptomycin, EGF (20 ng/mL), and FGF (20 ng/mL; Preprotech), Heparin 2 µg/mL (Sigma H33930) and N-2-supplement 1/100 (Gibco 17502-048). The animal ethics committee at the Ministry of Higher Education and Research in Paris approved all protocols involving live mice (protocol #17503 and #26137). C57BL/6 mice were purchased from Charles River and were given a week of acclimation before starting any experiment.

GL261 was transduced with a luciferase/mKate2 vector as described before[23]. GL261-luciferase and Nfpp10-luciferase cells (1.4 x 10^5 cells/2µL) were inoculated into the right caudate nucleus-putamen (AP +10 mm, DV +0.25 mm, ML +0.15 mm) of 7-8 weeks old C57BL/6 females using a stereotactic injection frame (David Kopf Instruments Tujunga, CA). Mice were

imaged using the IVIS Spectrum (PerkinElmer) 10 minutes following a 2 mg subcutaneous injection of luciferin (Sigma, L9504). The growth of GL261-luciferase and Nfpp10-luciferase cells was confirmed by two IVIS imaging one week apart of intracranial cell injection. We observed that mice with bioluminescence values lower than 5.00+E05 photon/second would not develop GBM tumors during the characterization of tumor growth in our mouse models. Therefore, we have included mice with bioluminescence values over 5.00+E05 photon/second. Mice were randomly placed into treatment arms once they passed the bioluminescence cutoff value.

Animals were treated with 200 µg of anti-CTLA-4 (Bristol-Myers Squibb, G1-XAS-Ab), anti-PD-L1 (Genentech, 6E11), and IgG1 (BXCELL, BE0083) and InVivoPure pH 6.5 Dilution Buffer (BXCELL, IP0065) for four doses. Unless stated otherwise, animals were sacrificed when they showed signs of tumor-associated illness (20% body weight loss or changes in behavior or posture).

Low-intensity pulsed ultrasound preclinical device

The pre-clinical ultrasound system (CarThera, Paris, France) was identical to that described in other studies[24] and shown in Fig. 2D. The system consisted of a 1 MHz, 10-mm diameter acoustic transducer that was coupled to the head of the mouse at a distance of 15-mm from the transducer. Sonications were performed for 120 seconds using a 25,000 cycle burst at a 1 Hz pulse repetition frequency and an acoustic pressure of 0.3 MPa as measured in water.

Ultrasound-mediated blood-brain barrier opening (UMBO)

UMBO was delivered to both UMBO and UMBO plus anti-PD-L1 groups. Anti-PD-L1 (6E11 Genentech) was administered intraperitoneally in all experiments at a dose of 200 µg sixty minutes before UMBO application. Mice were maintained under anesthesia with isoflurane (2%, 2L/min O_2). For each UMBO application, 10 ml/kg SonoVue® was injected through the retro-orbital route less than 10 seconds before the start of the ultrasound application. For each session, UMBO was validated using an additional control mouse. Each control mouse was injected intravenously with a solution of 2.7% Evans blue (Sigma, E2129) in phosphate buffer saline (PBS) at a dose of 4 mL/kg ten minutes post-sonication. All mice received 10 mL/kg warm saline injection in each treatment protocol before anesthesia to prevent any possible hypothermic effect. Intraperitoneal injection of anti-PD-L1 injection was given 60 min before sonication to ensure anti-PD-L1 absorption. UMBO test mice were sacrificed 15 minutes following Evans' blue injection, and their brain was harvested. The passage of Evans blue was assessed both visually and by ZEISS Axio-Scan fluorescence imaging of cryo-sectioned brains.

Pharmacokinetic (PK) analysis of therapeutic antibodies with and without UMBO

The PK analysis was performed using an identical molecular weight with similar conformational structure IgG1 isoform. Thirty-six mice were used in the pharmacokinetic experiment. Mice were separated in control and UMBO groups. Six-time points were selected as follows: 0.15, 0.3, 3, 6, 24, 48 and

96 hours. Each mouse received a 200 µg of nivolumab (Bristol-Meyers Squibb, New York, NY, USA) intravenous injection 10 minutes following the BBB opening. 100 µL of blood was collected through cardiac puncture using a pre-heparinized syringe. The serum was collected by centrifugation of the blood at 3500 rpm for 10 minutes.

All samples (plasma and brain) were then analyzed using ultra-performance liquid chromatography (UPLC) system coupled to mass spectrometry (LC-MS/ MS; MS-8060, Shimadzu, Japan). Peak integration and quantification were performed using LabSolutions Insight LC-MS software. Nivolumab was quantified with signature peptide ASGGITFSNSGMHWVR by nSMOL (Shimadzu, Japan)[25].

MRI data acquisition

Two GL261-bearing mice were used in the experiment. Two sessions per mouse were completed in 2 consecutive days to decrease any distress effect of long isoflurane exposure. MRI acquisitions were performed using a preclinical 11.7 T MRI scanner (Biospec, Bruker BioSpin, Germany) equipped with a CryoProbe dedicated to mouse brain imaging (Biospec, Bruker BioSpin, Germany). The total MRI experiment time was approximately 80 min per mouse (including MRI settings, acquisitions, and gadolinium injection), during which the animals were anesthetized with 1% isoflurane in O2 (2 L/min). Respiratory rate and body temperature were monitored while mice were restrained. For each animal, the protocol consisted in: (i) acquiring pre-gadolinium enhancement anatomical T_1-weighted (T_1w) images using a Multi-

Slice Multi Echo (MSME) sequence with the following parameters: T.R. = 400 ms, T.E. = 5 ms (one single echo), four averages, 14 slices, and resolution = 60x60x500 μm^3, (ii) following injection of a total volume of 100 µL of gadolinium (DOTAREM®,Guerbet, Aulnay-sous-Bois, France) at 0.5 mM and at physiological temperature in the tail vein of the mouse outside the MRI scanner, (iii) acquiring post-gadolinium T_1w images using the same sequence as used for (i) and, (iv) acquiring post-gadolinium injection T_2*-weighted (T_2*w) images using a multi gradient echo (MGE) sequence. MGE sequence was acquired with the following parameters: T.R.= 80 ms, ten echoes ranging from T.E. = 2.7 ms to 35.1 ms (echo spacing = 3.6 ms), and isotropic resolution of 60x60x60 μm^3. Gd enhancement volume was estimated on T_1w MRI pre and post BBB opening. On each T_1w, the hyper-intensity area corresponding to the Gd enhancement was segmented manually on FSLeyes. The volume of the Gd enhancement was measured as the total number of voxels multiplied by the spatial resolution and reflects the extent of the BBB opening.

mRNA sequencing

Six mice with a confirmed tumor of comparable sizes (as measured by bioluminescence imaging) were included in this experiment. Mice were divided into two groups (UMBO group and vehicle group). The vehicle group was treated with inVivoPure pH 6.5 Dilution Buffer (BXCELL, IP0065). Two treatment sessions (days 21 and 24) were applied in this experiment. Mice were sacrificed 24 hours after the last treatment by cervical dislocation, and

the right hemisphere was stored in 5 mL RNALater (Thermofisher AM7020). Lysing Matrix D (MBio, 6913050) was used to homogenize the collected brain tissues. mRNA was extracted using Maxwell RSC simply RNA automated RNA purification kit (Promega, AS1340). RNA quality was analyzed using high-sensitivity RNA chips (TapeStation). For RNA sequencing, NovaSeq 6000 sequencer (200 cycles, 800 million reads) and reagent kit were used. The reads (202 bp length, 100 million input reads) were mapped with the STAR v2.7.2a (default parameters) software to the reference genome (version GRCm38) on new junctions and known annotations. Mapping parameters were obtained from STAR outputs obtaining around 90% of unique mapped reads for all samples. Read counts from STAR were used as input for differential expression analysis using DESeq2. Furthermore, normalized counts were obtained using the variance-stabilizing transformation (VST) method from DESeq2 to be used as input for gene set variation analysis (GSVA)[26] to evaluate signature enrichment of microglia expression (Slc2a5, Siglech, P2ry12, Gpr34, P2ry13, Olfml3, Tmem119, Fcrls)[27], microglia sensome (96 genes)[28] or antigen presentation related genes (Ciita, Psme2b, Erap1, Irf1, Tapbp, Psme2, Psme1, Pdia3, Psme3, Tap1, B2m, Calr, Tap2, Hspa1a, H2-Ab1k, H2-K1, H2-D1)[29]. For heatmaps representation (ComplexHeatmap R package), VST gene expression values were first quantile normalized and log2 transformed; then converted to Z-scores by substracting the average expression value of gene i (Gi) of all samples from the gene expression the same Gi within sample x(Sx), the resulting value was

divided by the SD of Gi, the formula is: Z-scoreGiSx = (Expression GiSx-μGi)/σGi

Immunohistochemistry (IHC)

A 150 KDa rat IgG2 antibody targeting PD-L1 was used in our IHC staining (BXCELL, #BE0101). A Goat anti-rat secondary IgG (H+L) antibody (BA-9400) was used to detect the anti-PD-L1. Iba1 protein was detected using 1:1000 (Abcam, #ab178846). Mouse brains were fixed overnight in 4% paraformaldehyde (PFA), then immersed in 30% sucrose overnight for cryoprotection. Next, brains were stored in Tissue-Tek® O.C.T and stored at -80˚ Celsius. 10 μm cryosections were harvested using Leica CM1950 cryostat. Slides were stored at -80°C until analysis.

Quantitative digital droplet polymerase chain reaction (ddPCR)

GL261 tumor-bearing mice four weeks following cell inoculation were used in the ddPCR experiment. A single UMBO treatment was completed, and 30 minutes later, blood (100 μL) was collected in heparinized tubes through cardiac puncture. Whole blood DNA was extracted automatically using Maxwell® Blood DNA Purification Kit (AS1010). QX200 ddPCR EvaGeen® was utilized to detect *mKate2* and *Luciferase* genes in the extracted DNA. Primer3Plus web interface was used to design *mKate2*, and *Luciferase* primers and primers were purchased from Life Technologies. The following forward (FR) and reverse (RV) primers were used: luciferase-FR, TCCACGATGAAGAAGTGCTC; luciferase-RV, AGGCTACAAACGCTCTCATC;

mKate2-FR, GGTGAGCGAGCTGATTAAGG; and mKate2-RV, GGGTGTGGTTGATGAAGGTT.

Flow cytometry

Twenty mice with a confirmed tumor of comparable sizes were included in this experiment. Mice were separated into four groups: UMBO group, anti-PD-L1 (Genentech, 6E11) group, UMBO plus anti-PD-L1, and vehicle group (n=5/group). One treatment session was delivered in this experiment. Mice were perfused using cold distilled phosphate buffer saline (DPBS) ~16 hours after treatment. Brains were isolated immediately and stored in 2 mL ice-cold Hanks' balanced salt solution (HBSS). According to the manufacturer's protocol, the right hemisphere was isolated and mixed in the enzyme mix solution from the adult brain dissociation kit (Miltenyi Biotec, #130-107-677). Cells gentleMACS® Octo Dissociator with Heaters (#130-096-427) and gentleMACS C Tubes (#130-093-237) were used to perform mice brain dissociation. The number of dissociated cells was calculated using Scepter® 3.0 Handheld Cell Counter.

Samples were acquired on a spectral flow cytometer (Aurora, Cytek) and analyzed by FlowJo software (FlowJo, LLC). Briefly, cells were selected based on their morphology, doublets, and dead cells were excluded using (Biolegend, #423107) while tumor cells were excluded based on their *mKate* expression. Monocytes ($Ly6C^+$ $Ly6G^-$) and neutrophils ($Ly6C^+$ $Ly6G^+$) were excluded from non-tumoral live cells using Ly-6C (Biolegend, #128036) and Ly-6G (Biolegend, #127617). Microglia were identified based on their

expression of $CD11b^{+}$ and $CD45^{low}$ using CD45 (Biolegend, #103131) and CD11b (Biolegend, #101255). Activated microglia were identified as $CD68^{+}$ using (Biolegend, #137003). F4/80 marker (Biolegend, #123117) was used to determine macrophages in the $CD45^{high}$ $CD11b^{+}$ cell population. CD206 marker (Biolegend, #141729) was used to distinguish between subpopulations of macrophages. Lymphocytes $CD4^{+}$ (Biolegend, #100541) and $CD8^{+}$ (Biolegend, #100737) were identified on the $CD45^{+}$ $CD11b^{-}$ fraction of non-tumoral live cells. The percentage of each subpopulation was calculated and using in our flow cytometry analyses.

Statistical tests

Statistical analysis was performed using Prism software (GraphPad). Data are shown as mean values plus and minus standard error of the mean (SEM). Statistical significance of differences between groups was verified using appropriate statistical tests. Significance level were denoted with asterisks: * for $p \leq 0.05$; ** for $p \leq 0.01$; *** for $p \leq 0.001$, and **** for $p \leq$ 0,0001.

Results

Selection of GBM mouse model and immunotherapeutic antibody to combine with UMBO

Pilot studies with no UMBO using both GL261-luciferase and Nfpp10-luciferase orthotopic GBM mouse models were performed. These two experiments aimed to determine the effect of anti-PD-L1 and anti-CTLA-4 in our GBM mouse model and select the best candidates to combine with UMBO.

Anti-PD-L1 antibody alone has shown an early regression in tumor growth (Figure 1-A) and shown a limited effect on survival of GL261-bearing (Figure 1-B), Anti-CTLA-4 treatment did not affect tumor growth (Figure 1-A) or animal survival (Figure 1-B). No treatments had an impact on mouse body weight (Figure 1-C).

Interestingly, anti-PD-L1 antibody showed better efficacy in the Nfpp10 GBM mouse model than GL261-bearing mice (Figure 1-E). Anti-PD-L1 treatment did not reduce tumor growth (Figure 1-D) yet did increase the number of long-term survivors (3/6) (Figure 1-E).

We evaluated BBB integrity in both GBM mouse models. Assessment of BBB disruption was performed using 1.2 mg of Hoechst 33342 (Sigma)

diluted in PBS and injected intravenously 20 min prior to sacrifice. Hoechst staining was not detected in normal brain tissue (Figure 2-A), yet higher staining intensity was observed in brain tissue harvested from Nfpp10-bearing mice compared to GL261-bearing mice. Furthermore, using RT-PCR we evaluated the quantitative expression of PD-L1 in GL261 and Nfpp10 cell lines. PD-L1 expression was significantly higher in the GL261 cell line compared to Nfpp10 (Figure 4-D). Those as mentioned earlier indicate a higher BBB permeability and may explain higher efficacy of anti-PD-L1 therapeutics antibodies in the Nfpp10 GBM mouse model. Additionally, the BBB integrity is higher in GL261-bearing mice. Overall, this makes the anti-PD-L1 antibody the best candidate to combine UMBO in the GL261 GBM mouse model.

Repeated UMBO is safe and effective in immunocompetent mice

UMBO parameters were previously optimized in our setting using athymic nude mice[30], we evaluated UMBO parameters and treatment frequency in GL261 GBM mouse models. T_1w MRI also (Figure 2-E) showed a marked gadolinium contrast enhancement within an hour following the ultrasound emission (Figure 2-E).

Biweekly UMBO (4 sonications in total without drug) was evaluated in the GL261-bearing mice. Mouse weight was unaffected (Figure 3-B) and no significant difference in the overall survival (OS) between UMBO and non-treated groups were observed (Figure 3-A). Overall, the UMBO parameters

used for repeated BBB opening were safe and well-tolerated in GL261-bearing mice.

Using bulk RNA sequencing, we attempted to check whether UMBO modulates antigen presentation related genes compared to the vehicle group. We observed that UMBO did not influence antigen presentation (Figure 3-C) or affect microglia gene expression (Figure 3-D). Microglial ability to sense changes in the cellular environment was recently termed as microglia sensome[28]. We attempted to use the same gene signature to evaluate microglial sensome with and without UMBO. Interestingly, UMBO significantly induced the expression of gene signatures for microglial sensome (Figure-3E).

Additionally, we aimed to evaluate whether UMBO could enhance the leakage of circulating tumor DNA to the bloodstream. GL261-bearing mice four weeks following GL261 cell grafting were used in the experiment. We observed a significant elevation in the number of copies for both *luciferase* (Figure 5-D) and *mKate2* (Figure 5-E) in the UMBO treated group compared to the control.

UMBO dramatically increased the efficacy of anti-PD-L1 in GL261-bearing mice

We next investigated the combined effect of UMBO combined with anti-PD-L1 in the GL261 GBM mouse model. Mice with comparable bioluminescence values were divided into five groups: (i) UMBO group, (ii)

anti-PD-L1 group, (iii) UMBO plus anti-PD-L1 group, (iv) IgG1 group, and (v) IgG1 plus UMBO group.

We have not observed any toxic effect in UMBO plus anti-PD-L1 treated mice versus control mice (Figure 4-A). UMBO and anti-PD-L1 reduced tumor growth (Figure 4-B,E). Interestingly, mice who received an anti-PD-L1 antibody with UMBO showed a (13/17) 76 % long-term survivors (over 100 days) compared to (4/15) 26 % in anti-PD-L1 alone and (0/16) 0% in control groups. Kaplan-Meier estimate shows a significant difference in UMBO's overall survival plus anti-PD-L1 treated mice versus anti-PD-L1 alone treated mice (Figure 4-C). Furthermore, a higher significance difference (Figure 4-C) was observed in UMBO plus anti-PD-L1 treated mice compared to the IgG1 plus UMBO treated mice in the GL261 GBM mouse model.

UMBO increased the penetration of anti-PD1 and anti-PD-L1 antibodies into the brain parenchyma

The BBB blocks large therapeutic agents such as antibodies. With UMBO, we attempted to measure delivery of antibodies to the brain parenchyma. IHC staining of anti-PD-L1 (BXCELL, BE0101) confirmed UMBO's ability to deliver anti-PD-L1 to the right hemisphere brain parenchyma (Figure 5-A). Furthermore, an already clinically optimized nSMOL[25] method was used to compare a size-matched IgG1 antibody’s pharmacokinetics with and without UMBO. Three C57BL/6 mice per time point (six-time points) per group were used in the analysis. We observed a comparable serum concentration of nivolumab in control and UMBO-treated mice.

Interestingly, higher concentrations of nivolumab were detected in mice brains treated with nivolumab plus UMBO. As expected, we detected a negligible concentration (≤0.2 µg/200mg brain) of nivolumab in control mice brains (Figure 5-C). The maximum concentration (Cmax) of nivolumab in normal brain tissue was detected at 24 hours and started to decline and reach a negligeable concentration at 96 hours. Therefore, a regimen of a biweekly antibody administration was performed. An analysis of the ratio changes in brain to plasma concentration shows that UMBO enhanced the ratio of nivolumab passage across the BBB at 3, 24, 48 hours but not at 96 hours (Figure 5-B).

UMBO plus anti-PD-L1 activates microglia and modulates microglial phenotype

In order to further explore the potentialities of anti-PD-L1 efficacy with UMBO, we studied the presence of the different immune populations in our treatment groups by flow cytometry. Interestingly, we found that UMBO plus anti-PD-L1 significantly enhanced the percentage of activated microglia compared to anti-PD-L1 treatment alone (Figure 6-A). Additionally, although UMBO alone was not associated with a significant enhancement of activated microglia percentage compared to the vehicle group, a trend was observed (p=0.150). On the other hand, we did not observe any significant changes in the percentage of $CD8^+$ and $CD4^+$ T-lymphocytes or $CD206^+$ macrophages in all groups. Immunofluorescence staining of microglia in anti-PD-L1 plus UMBO treated group confirmed this finding and showed a phenotype of activated

microglia. In addition, using IHC we observed a double nucleus staining of Iba1 in the UMBO plus anti-PD-L1 treated GL261-bearing mice suggests a possible induction of microglia cell division (Figure 6-G).

Discussion

UMBO and several innovative strategies continuously evolve to overcome the BBB by increasing drug delivery[8]. Immunotherapies, including ICIs and cell therapies, have revolutionized multiple solid tumors' treatments through activating the general antitumor immune response. The CheckMate-143 phase 3 clinical trial was initiated to evaluate the effect of nivolumab *versus* bevacizumab. Unfortunately, nivolumab did not demonstrate higher efficacy compared to bevacizumab, which has been shown to have no effect on OS in GBM patients. Several reasons might explain the low efficacy of ICIs in GBM: (i) low tumor mutation load, (ii) lack of predictor of response and lack of selection of patients, (iii) low penetration of ICIs within the brain parenchyma, (iv) low peripheral priming, (v) local immunosuppression and (vi) low penetration of T-lymphocytes[31].

We explored the BBB as the limitation for antibody and lymphocytes penetration and priming and attempted to evaluate UMBO's effect on the penetrating large therapeutics to the brain and modulating the immune microenvironment in GBM mouse model. Our data confirmed the limited efficacy of ICIs efficacy in the Gl261-bearing and Nfpp10-bearing mouse models. Consistent with our data, Reardon et al. showed limited efficacy of

anti-PD-L1 and anti-CTLA-4 in GL261-bearing mice model although they used a different treatment regimen[32].

To the best of our knowledge, this is the first research article that report a dramatic increase in the overall survival of GL261-bearing mice when treated with UMBO plus anti-PD-L1. Indeed, 76% of GL261-bearing mice treated with anti-PD-L1 plus UMBO survive longer than 100 days compared to 26% for GL261 mice treated with anti-PD-L1 alone. Next, we tried to understand the mechanisms involved in the anti-tumor effect. We initially hypothesized that the BBB was responsible for the limited efficacy by blocking anti-CTLA-4 and anti-PD-L1 from reaching the GBM tumor. This hypothesis is consistent with a recent study that reported an enhanced efficacy of ICIs following their delivery to brain tumors[33]. Recently, it was reported that focused ultrasound enhanced the delivery of intranasal administration of anti-PD-L1 but not overall survival of GL261-bearning mice[34].

In our setting, we reported that UMBO enhanced antibody concentration up to 28-fold compared to control. UMBO was optimized to disturb one hemisphere; however, in our PK analysis, we used a whole-brain homogenization method; therefore, local concentrations of nivolumab could have been even higher. Consistent with our data, a study has shown that UMBO enhanced the delivery of bevacizumab ~149 KDa to the brain parenchyma by 5.7 to 56.7 folds compared to non-sonicated brain in a glioma mouse model[35]. UMBO plus 200 µg of the anti-PD-L1 biweekly treatment

regimen was used to maintain the higher concentration of anti-PD-L1 within the brain parenchyma. Immune checkpoint blockade with anti-PD-L1 was performed on day 14 post-inoculation to allow for T-lymphocytes depletion[36]

GBM tumors have low chances of extracranial metastases with negligible risk for GBM spreading after surgical brain biopsies. UMBO stimulates a detectable peripheral circulation of GL261 DNA[37]. Zhu et al., published article investigated the possibility of using UMBO for liquid biopsies in GBM models. They observed a detectable level of green fluorescent protein mRNA 20-mins following UMBO in the GL261-GFP expressing mouse model[38] supports the passage of tumor material from the brain to blood flow stream.

The priming effect of circulating DNA could activate naïve T-lymphocytes through their exposure to new antigens. As mentioned previously, the BBB protects the tumor from T-lymphocytes infiltration and immune activation. Thus, detecting GL261 tumors in the peripheral circulation might activate the global antitumor effect. Further functional demonstration of lymphocyte activation should be performed to evaluate any priming effect of UMBO.

Our results showing microglia activation in the UMBO plus anti-PD-L1 treated GL261-bearing mice suggest a possible mechanism for the observed enhanced therapeutic efficacy of anti-PD-L1. Our flow cytometry analysis is consistent with a newly published article that observed a higher ratio of Iba-1 staining in sonicated brain regions compared to non-sonicated regions. However, this difference was not statistically significant[39]. PD-L1 is expressed on the cell surface of both GL261 and microglia[40]. A possible effect on

microglia phenotype might be related to the combined effect of UMBO and anti-PD-L1 delivery to the brain parenchyma. Activated microglia might have an impact on the cytotoxic effect against GL261 tumor cells[41]. Therefore, a further investigation of the combined effect of UMBO and immune checkpoint inhibitors on microglia phenotype should be addressed.

To date, there is no clear evidence on the effect of UMBO on T-lymphocytes passage to the brain. We have not observed any significant elevation in the percentage of $CD8^+$ and $CD4^+$ T-lymphocytes at one timepoint (~16 hours). This effect might be related to the timing of sample collection. We have not evaluated the effect of our treatment regimen at later time points. We have seen a delayed antitumor effect in UMBO and anti-PD-L1 group which could be related to a delayed effect on T-lymphocytes. Furthermore, we have not analyzed any subpopulations of $CD8^+$ T lymphocytes *i.e.,* PD-1^+ $CD8^+$ T-lymphocytes.

Syngeneic mice models and especially the GL261 mouse model used in our experiments is one limitation of the current study. The GL261 mouse model : (i) has a high mutation load which is not consistent with GBM patients and (ii) a variability in terms of responses to ICIs *in vivo*[36]. Another limitation of our findings is the inability to demonstrate functional analysis of the role of UMBO in priming naïve T-lymphocytes through their exposure to new antigens. Additional functional analysis on the effect of UMBO plus anti-PD-L1 would explain the dramatic effect on OS that was observed in our study.

Conclusions

Our study showed statistically significant increased brain penetration and efficacy of anti-PD-L1 in GL261-bearing mice when delivered by UMBO. We have also provided clear evidence of the possible safe and effective delivery of large therapeutic agents using UMBO. Further investigations are needed to confirm the impact of UMBO on brain penetration and efficacy of chemotherapeutic agents and anti-PD-L1 to overcome the resistance of GBM to the current treatments.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement #766069 (GLIO-TRAIN). We also acknowledge Genentech for providing anti-PD-L1 (Clone 6E11) antibody. Additionally, we appreciate our valuable discussions with Dr. Michel Mallat and his comments regarding microglia. Finally, we acknowledge the Salk institute for biological studies, CA 92037 USA for providing us with the Nfpp10 cells.

The research leading to these results has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06. Institut Universitaire de Cancérologie. INCA-DGOS-Inserm_12560 SiRIC CURAMUS is financially supported by the French National Cancer Institute, the French Ministry of Solidarity and Health and Inserm.

Conflicts of interest

AI reports grants and travel funding from CarThera, research grants from Transgene, Sanofi, Air Liquide and, Nutritheragene, travel funding from Leo Pharma, advisory board from Novocure and Leo Pharma. AC is a paid consultant of CarThera and have ownership interest in CarThera. All other authors have no conflict of interests. CS, GB, MC are employees of CarThera.

Author contribution

MA designed and performed experiments, analyzed data, and wrote the paper. MA, NL, EQ, CG, LG, NZ, LM, AP performed the experiments. IH.; performed bioinformatic analyses of mRNA sequencing data. AI supervised the research.

References

1. Ostrom QT, Bauchet L, Davis FG, et al. The epidemiology of glioma in adults: a "state of the science" review. *Neuro Oncol.* 2014; 16(7):896-913.
2. Pace A, Dirven L, Koekkoek JAF, et al. European Association for Neuro-Oncology (EANO) guidelines for palliative care in adults with glioma. *The Lancet. Oncology.* 2017; 18(6):e330-e340.
3. Atiq A, Parhar I. Anti-neoplastic Potential of Flavonoids and Polysaccharide Phytochemicals in Glioblastoma. *Molecules (Basel, Switzerland).* 2020; 25(21).
4. Touat M, Idbaih A, Sanson M, Ligon KL. Glioblastoma targeted therapy: updated approaches from recent biological insights. *Annals of oncology : official journal of the European Society for Medical Oncology.* 2017; 28(7):1457-1472.
5. Weenink B, French PJ, Sillevis Smitt PAE, Debets R, Geurts M. Immunotherapy in Glioblastoma: Current Shortcomings and Future Perspectives. *Cancers.* 2020; 12(3).
6. Wang N, Jain RK, Batchelor TT. New Directions in Anti-Angiogenic Therapy for Glioblastoma. *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics.* 2017; 14(2):321-332.
7. Lara-Velazquez M, Al-Kharboosh R, Jeanneret S, et al. Advances in Brain Tumor Surgery for Glioblastoma in Adults. *Brain sciences.* 2017; 7(12).
8. Drean A, Goldwirt L, Verreault M, et al. Blood-brain barrier, cytotoxic chemotherapies and glioblastoma. *Expert review of neurotherapeutics.* 2016; 16(11):1285-1300.
9. Zhang DY, Dmello C, Chen L, et al. Ultrasound-mediated Delivery of Paclitaxel for Glioma: A Comparative Study of Distribution, Toxicity, and Efficacy of

Albumin-bound Versus Cremophor Formulations. *Clinical cancer research : an official journal of the American Association for Cancer Research.* 2020; 26(2):477-486.

10. Idbaih A, Canney M, Belin L, et al. Safety and Feasibility of Repeated and Transient Blood–Brain Barrier Disruption by Pulsed Ultrasound in Patients with Recurrent Glioblastoma. 2019; 25(13):3793-3801.

11. Sheikov N, McDannold N, Vykhodtseva N, Jolesz F, Hynynen K. Cellular mechanisms of the blood-brain barrier opening induced by ultrasound in presence of microbubbles. *Ultrasound in medicine & biology.* 2004; 30(7):979-989.

12. Carpentier A, Canney M, Vignot A, et al. Clinical trial of blood-brain barrier disruption by pulsed ultrasound. *Sci Transl Med.* 2016; 8(343).

13. Martínez-Fernández R, Máñez-Miró JU, Rodríguez-Rojas R, et al. Randomized Trial of Focused Ultrasound Subthalamotomy for Parkinson's Disease. *The New England journal of medicine.* 2020; 383(26):2501-2513.

14. Idbaih A, Canney M. Safety and Feasibility of Repeated and Transient Blood-Brain Barrier Disruption by Pulsed Ultrasound in Patients with Recurrent Glioblastoma. 2019; 25(13):3793-3801.

15. Rezai AR, Ranjan M, D'Haese P-F, et al. Noninvasive hippocampal blood−brain barrier opening in Alzheimer's disease with focused ultrasound. *Proceedings of the National Academy of Sciences.* 2020; 117(17):9180-9182.

16. Tarhini AA. Tremelimumab: a review of development to date in solid tumors. *Immunotherapy.* 2013; 5(3):215-229.

17. Hsu FS, Su CH, Huang KH. A Comprehensive Review of US FDA-Approved Immune Checkpoint Inhibitors in Urothelial Carcinoma. *Journal of immunology research.* 2017; 2017:6940546.

18. Hao C, Chen G, Zhao H, et al. PD-L1 Expression in Glioblastoma, the Clinical and Prognostic Significance: A Systematic Literature Review and Meta-Analysis. *Frontiers in oncology.* 2020; 10:1015.

19. Azoury SC, Straughan DM, Shukla V. Immune Checkpoint Inhibitors for Cancer Therapy: Clinical Efficacy and Safety. *Curr Cancer Drug Targets.* 2015; 15(6):452-462.

20. Reardon DA, Brandes AA, Omuro A, et al. Effect of Nivolumab vs Bevacizumab in Patients With Recurrent Glioblastoma: The CheckMate 143 Phase 3 Randomized Clinical Trial. *JAMA oncology.* 2020; 6(7):1003-1010.

21. Awada G, Ben Salama L, De Cremer J, et al. Axitinib plus avelumab in the treatment of recurrent glioblastoma: a stratified, open-label, single-center phase 2 clinical trial (GliAvAx). *Journal for immunotherapy of cancer.* 2020; 8(2).

22. Friedmann-Morvinski D, Bushong EA, Ke E, et al. Dedifferentiation of Neurons and Astrocytes by Oncogenes Can Induce Gliomas in Mice. 2012; 338(6110):1080-1084.

23. Plessier A, Le Dret L, Varlet P, et al. New in vivo avatars of diffuse intrinsic pontine gliomas (DIPG) from stereotactic biopsies performed at diagnosis. *Oncotarget.* 2017; 8(32):52543-52559.

24. Zhang DY, Dmello C, Chen L, et al. Ultrasound-mediated Delivery of Paclitaxel for Glioma: A Comparative Study of Distribution, Toxicity, and Efficacy of Albumin-bound Versus Cremophor Formulations. *Clinical Cancer Research.* 2020; 26(2):477-486.

25. Iwamoto N, Yokoyama K, Takanashi M, Yonezawa A, Matsubara K, Shimada T. Application of nSMOL coupled with LC-MS bioanalysis for monitoring the Fc-fusion biopharmaceuticals Etanercept and Abatacept in human serum. *Pharmacol Res Perspect.* 2018; 6(4):e00422.

26. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics.* 2013; 14(1):7.

27. Haage V, Semtner M, Vidal RO, et al. Comprehensive gene expression meta-analysis identifies signature genes that distinguish microglia from peripheral monocytes/macrophages in health and glioma. *Acta Neuropathologica Communications.* 2019; 7(1):20.

28. Maas SLN, Abels ER, Van De Haar LL, et al. Glioblastoma hijacks microglial gene expression to support tumor growth. *Journal of neuroinflammation.* 2020; 17(1):120.

29. Schmidt J, Smith AR, Magnin M, et al. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Reports Medicine.* 2021; 2(2):100194.

30. Dréan A, Lemaire N, Bouchoux G, et al. Temporary blood-brain barrier disruption by low intensity pulsed ultrasound increases carboplatin delivery and efficacy in preclinical models of glioblastoma. *Journal of neuro-oncology.* 2019; 144(1):33-41.

31. Beccaria K, Canney M, Bouchoux G, et al. Ultrasound-induced blood-brain barrier disruption for the treatment of gliomas and other primary CNS tumors. *Cancer letters.* 2020; 479:13-22.

32. Reardon DA, Gokhale PC, Klein SR, et al. Glioblastoma Eradication Following Immune Checkpoint Blockade in an Orthotopic, Immunocompetent Model. *Cancer Immunol Res.* 2016; 4(2):124-135.

33. Guo H, Wang R, Wang D, et al. Deliver anti-PD-L1 into brain by p-hydroxybenzoic acid to enhance immunotherapeutic effect for glioblastoma. *Journal of controlled release : official journal of the Controlled Release Society.* 2020; 320:63-72.

34. Ye D, Yuan J, Yue Y, Rubin JB, Chen H. Focused Ultrasound-Enhanced Delivery of Intranasally Administered Anti-Programmed Cell Death-Ligand 1 Antibody to an Intracranial Murine Glioma Model. *Pharmaceutics.* 2021; 13(2).

35. Liu HL, Hsu PH, Lin CY, et al. Focused Ultrasound Enhances Central Nervous System Delivery of Bevacizumab for Malignant Glioma Treatment. *Radiology.* 2016; 281(1):99-108.

36. Aslan K, Turco V, Blobner J, et al. Heterogeneity of response to immune checkpoint blockade in hypermutated experimental gliomas. *Nature Communications.* 2020; 11(1):931.

37. Lun M, Lok E, Gautam S, Wu E, Wong ET. The natural history of extracranial metastasis from glioblastoma multiforme. *Journal of neuro-oncology.* 2011; 105(2):261-273.

38. Zhu L, Cheng G, Ye D, et al. Focused Ultrasound-enabled Brain Tumor Liquid Biopsy. *Scientific Reports.* 2018; 8(1):6553.

39. Sinharay S, Tu T-W, Kovacs ZI, et al. In vivo imaging of sterile microglial activation in rat brain after disrupting the blood-brain barrier with pulsed focused ultrasound: [18F]DPA-714 PET study. *Journal of neuroinflammation.* 2019; 16(1):155.

40. Chen Q, Xu L, Du T, et al. Enhanced Expression of PD-L1 on Microglia After Surgical Brain Injury Exerts Self-Protection from Inflammation and Promotes Neurological Repair. *Neurochemical research.* 2019; 44(11):2470-2481.
41. Li Y, Zhang R, Hou X, et al. Microglia activation triggers oligodendrocyte precursor cells apoptosis via HSP60. *Mol Med Rep.* 2017; 16(1):603-608.

Figure 1

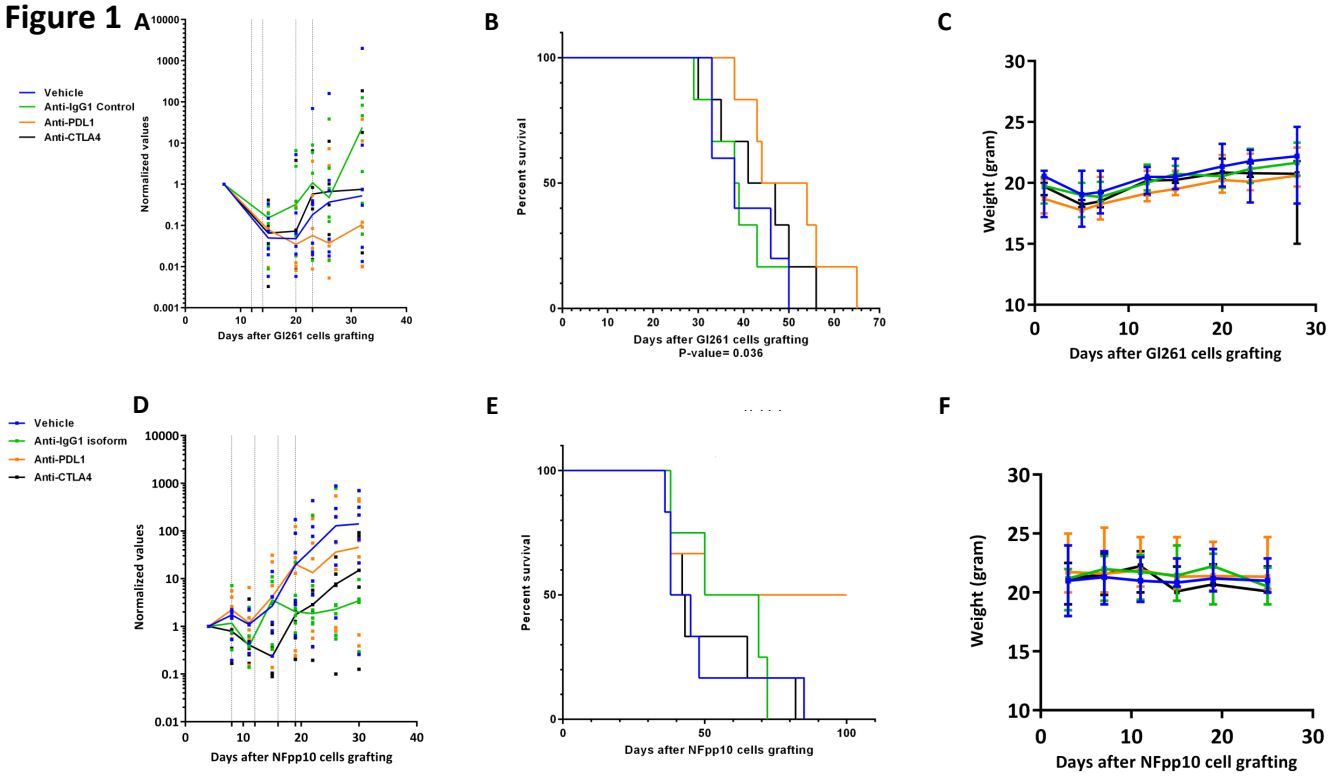

A
Vehicle
Anti-IgG1 Control
Anti-PDL1
Anti-CTLA4
Normalized values
Days after GI261 cells grafting
B
Percent survival
Days after GI261 cells grafting
P-value= 0.036
C
Weight (gram)
Days after GI261 cells grafting
D
Vehicle
Anti-IgG1 isoform
Anti-PDL1
Anti-CTLA4
Normalized values
Days after NFpp10 cells grafting
E
Percent survival
Days after NFpp10 cells grafting
F
Weight (gram)
Days after NFpp10 cell grafting

Figure 2

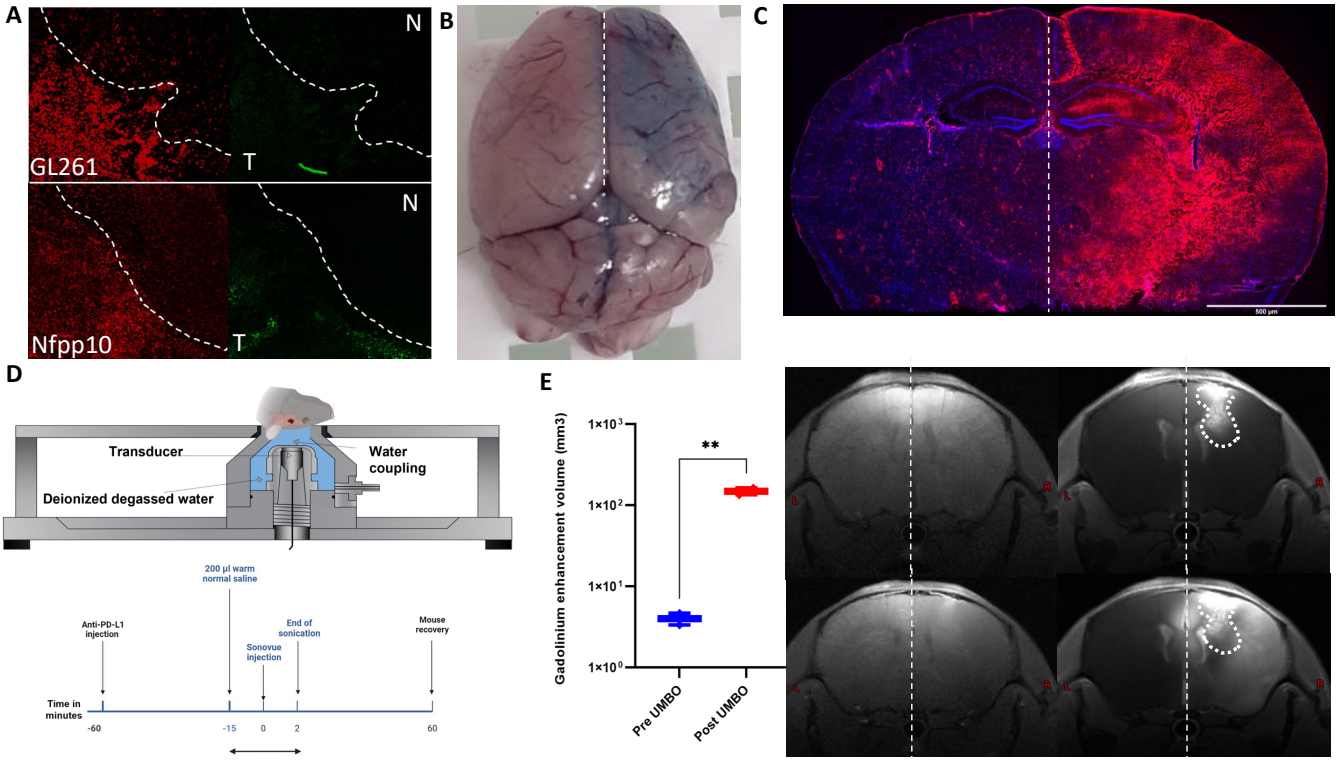

A
N
GL261
T
N
Nfpp10
T
B
C
D
Transducer
Water coupling
Deionized degassed water
200 µl warm normal saline
Anti-PD-L1 injection
Sonovue injection
End of sonication
Mouse recovery
Time in minutes
-60
-15
0
2
60
E
Gadolinium enhancement volume (mm3)
1×10^3
1×10^2
1×10^1
1×10^0
**
Pre UMBO
Post UMBO

Figure 3

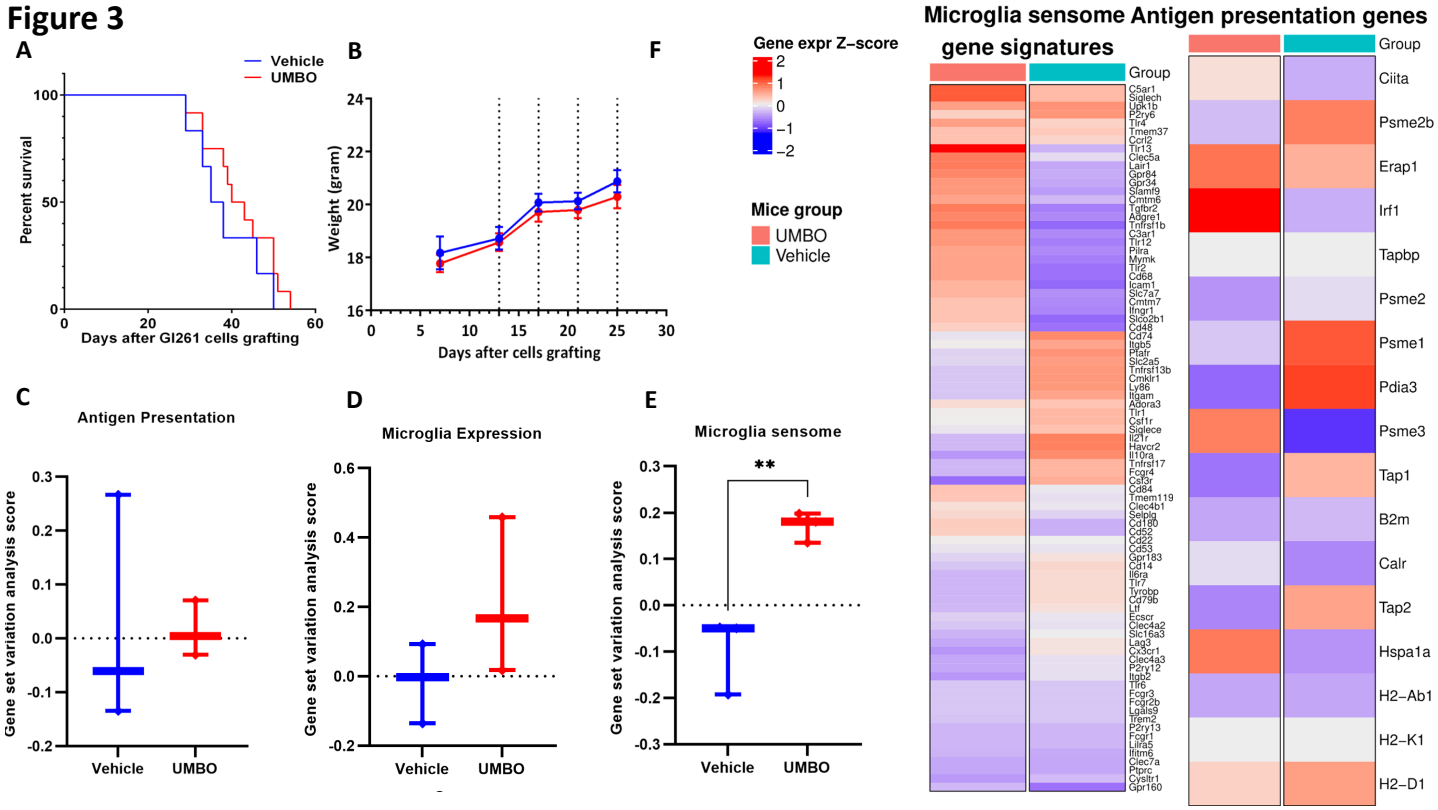

A
Vehicle
UMBO
Percent survival
Days after GI261 cells grafting
B
Weight (gram)
Days after cells grafting
F
Gene expr Z-score
Mice group
UMBO
Vehicle
C
Antigen Presentation
Gene set variation analysis score
Vehicle
UMBO
D
Microglia Expression
E
Microglia sensome
**
Microglia sensome Antigen presentation genes
gene signatures
Group
Ciita
Psme2b
Erap1
Irf1
Tapbp
Psme2
Psme1
Pdia3
Psme3
Tap1
B2m
Calr
Tap2
Hspa1a
H2–Ab1
H2–K1
H2–D1

Figure 4

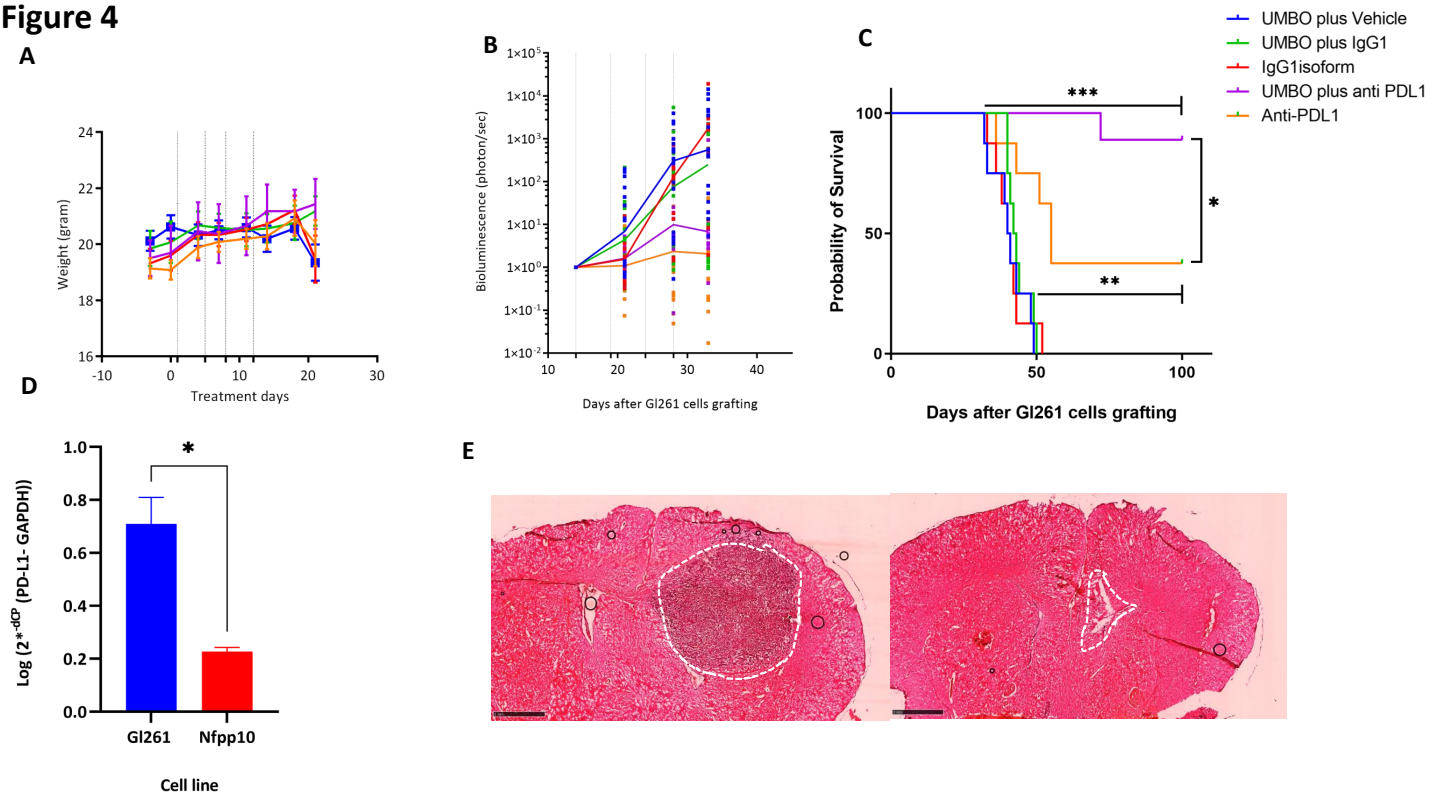

A
Weight (gram)
Treatment days
B
Bioluminescence (photon/sec)
Days after Gl261 cells grafting
C
UMBO plus Vehicle
UMBO plus IgG1
IgG1isoform
UMBO plus anti PDL1
Anti-PDL1
Probability of Survival
Days after Gl261 cells grafting
***
**
*
D
Log (2^-ddCp (PD-L1- GAPDH))
Gl261
Nfpp10
Cell line
E

Figure 5

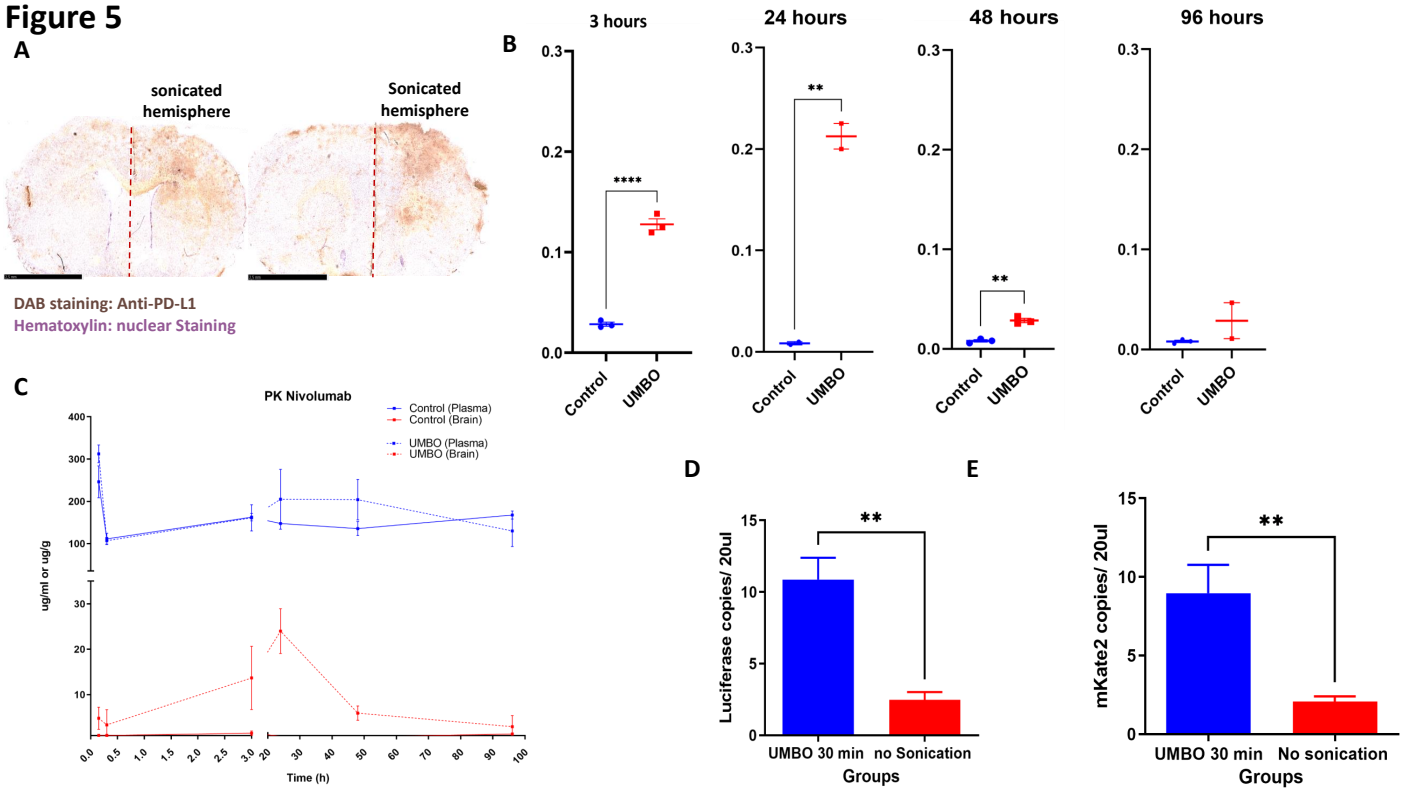
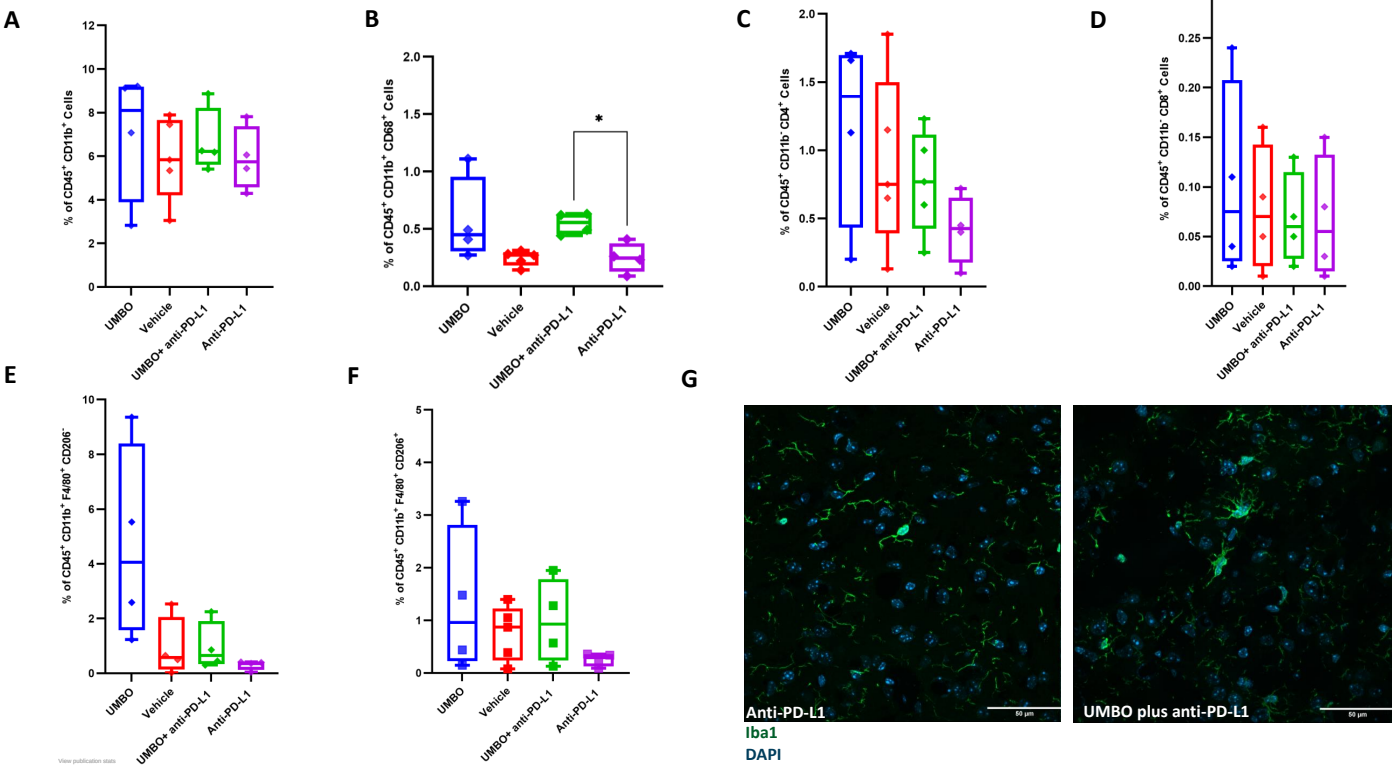

A
sonicated hemisphere
Sonicated hemisphere
DAB staining: Anti-PD-L1
Hematoxylin: nuclear Staining
B
3 hours
24 hours
48 hours
96 hours
****
**
**
Control
UMBO
C
PK Nivolumab
Control (Plasma)
Control (Brain)
UMBO (Plasma)
UMBO (Brain)
ug/ml or ug/g
Time (h)
D
Luciferase copies/ 20ul
**
UMBO 30 min
no Sonication
Groups
E
mKate2 copies/ 20ul
**
UMBO 30 min
No sonication
Groups

Figure 6


A
% of CD45+ CD11b+ Cells
B
% of CD45+ CD11b+ CD68+ Cells
*
C
% of CD45+ CD11b- CD4+ Cells
D
% of CD45+ CD11b- CD8+ Cells
E
% of CD45+ CD11b+ F4/80+ CD206-
F
% of CD45+ CD11b+ F4/80+ CD206+
UMBO
Vehicle
UMBO+ anti-PD-L1
Anti-PD-L1
G
Anti-PD-L1
UMBO plus anti-PD-L1
Iba1
DAPI

The Canadian Journal of Neurological Sciences (2022), 1–9
doi:10.1017/cjn.2022.5

Original Article

Expression and Prognostic Value of CD80 and CD86 in the Tumor Microenvironment of Newly Diagnosed Glioblastoma

Mohammed H. Ahmed[1,2], Isaias Hernández-Verdin[1,2], Franck Bielle[3], Maite Verreault[1], Julie Lerond[1], Agusti Alentorn[4], Marc Sanson[4] and Ahmed Idbaih[4]

Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital DE LA Pitié Salpêtrière, Paris, France, Faculté DE Médecine Paris-Sud, Université Paris-Saclay, Saint-Aubin, France, Sorbonne Université, Inserm, CNRS, UMR S 1127, Institut du Cerveau, ICM, AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière - Charles Foix, Service de Neuropathologie-Escourolle, Paris, France and Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, AP-HP, Hôpital DE LA Pitié Salpêtrière, DMU Neurosciences, Service de Neurologie 2-Mazarin, Paris, France

ABSTRACT: ***Background:*** Strategies to modulate the tumor microenvironment (TME) have opened new therapeutic avenues with dramatic yet heterogeneous intertumoral efficacy in multiple cancers, including glioblastomas (GBMs). Therefore, investigating molecular actors of TME may help understand the interactions between tumor cells and TME. Immune checkpoint proteins such as a Cluster of Differentiation 80 (CD80) and CD86 are expressed on the surface of tumor cells and infiltrative tumor lymphocytes. However, their expression and prognostic value in GBM microenvironment are still unclear. ***Methods:*** In this study, we investigated, in a retrospective local discovery cohort and a validation TCGA dataset, expression of CD80 and CD86 at mRNA level and their prognostic significance in response to standard of care. Furthermore, CD80 and CD86 at the protein level were investigated in the discovery cohort. ***Results:*** Both CD80 and CD86 are expressed heterogeneously in the TME at mRNA and protein levels. In a univariate analysis, the mRNA expression of CD80 and CD86 was not significantly correlated with OS in both local OncoNeuroTech dataset and TCGA datasets. CD80 and CD86 mRNA high expression was significantly associated with shorter progression free survival PFS ($p < 0.05$). These findings were validated using the TCGA cohort; higher CD80 and CD86 expressions were correlated with shorter PFS ($p < 0.05$). In multivariate analysis, CD86 mRNA expression was an independent prognostic factor for PFS in the TCGA dataset only ($p < 0.05$). ***Conclusion:*** CD86 could be used as a potential biomarker for the prognosis of GBM patients treated with immunotherapy; however, additional studies are needed to validate these findings.

Keywords: Glioblastoma; Microenvironment; Immune checkpoint proteins; Prognosis

(Received 13 November 2021; final revisions submitted 4 January 2022; date of acceptance 6 January 2022)

Introduction

Glioblastoma (GBM) is the most common and aggressive glioma in adults. The latest World Health Organization guideline classifies GBM as grade IV glioma.[1] Over the last years, massive efforts have led to a better understanding of the pathology and the genetic of GBM.[2] To date, the most effective and approved standard therapeutic regimen is maximum surgical resection of the tumor followed by concurrent chemoradiation and adjuvant chemotherapy with temozolomide.[1] Despite this very intensive therapeutic regimen, newly diagnosed GBM patients have a dismal outcome with a median overall survival (OS) below 18 months.[3] The main known prognostic factors are (i) age, (ii) Karnofsky performance status (KPS), (iii) *MGMT* promoter methylation status, and (iv) IDH mutational status.[4]

Immunotherapies have dramatically improved melanoma prognosis[5] and other nonneurological solid tumors.[5] In the setting of primary brain cancer, results from clinical trials are still disappointing.[6] Nonetheless, specific GBM patients responded, supporting the identification of biomarkers to stratify patients in the prescription of immunotherapies. Immune checkpoint proteins such as Cluster of Differentiation 80 (CD80; known as B7-1) and CD86 (known as B7-2) are expressed on the surface of tumor[7] and immune cells[8] but not glial cells.[9] CD80 protein expression was observed in infiltrative tumor lymphocytes in melanoma.[10]

Cytotoxic T-lymphocyte-associated antigen–4 (CTLA-4) and Cluster of Differentiation 28 (CD28) are located on T-lymphocytes. Both CD28 and CTLA-4 proteins bind to their ligands on the antigen-presenting cells and major histocompatibility complex.[11] The interaction between immune checkpoint proteins and their coreceptor at the surface of T-lymphocytes delivers the signal to activate or inhibit T cells function, that is, CTLA-4 has a higher affinity to CD80 and CD86, and when bound to its ligands, T cells remain exhausted.[12]

In preclinical studies, antibodies targeting CTLA-4 were used in preclinical studies to block CTLA-4 from binding to its ligands.[13]

Corresponding author: Ahmed Idbaih, Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, AP-HP, Hôpital DE LA Pitié Salpêtrière, DMU Neurosciences, Service de Neurologie 2-Mazarin, Paris F-75013, France. Email: ahmed.idbaih@aphp.fr

Cite this article: Ahmed MH, Hernández-Verdin I, Bielle F, Verreault M, Lerond J, Alentorn A, Sanson M, and Idbaih A. Expression and Prognostic Value of CD80 and CD86 in the Tumor Microenvironment of Newly Diagnosed Glioblastoma. *The Canadian Journal of Neurological Sciences* https://doi.org/10.1017/cjn.2022.5

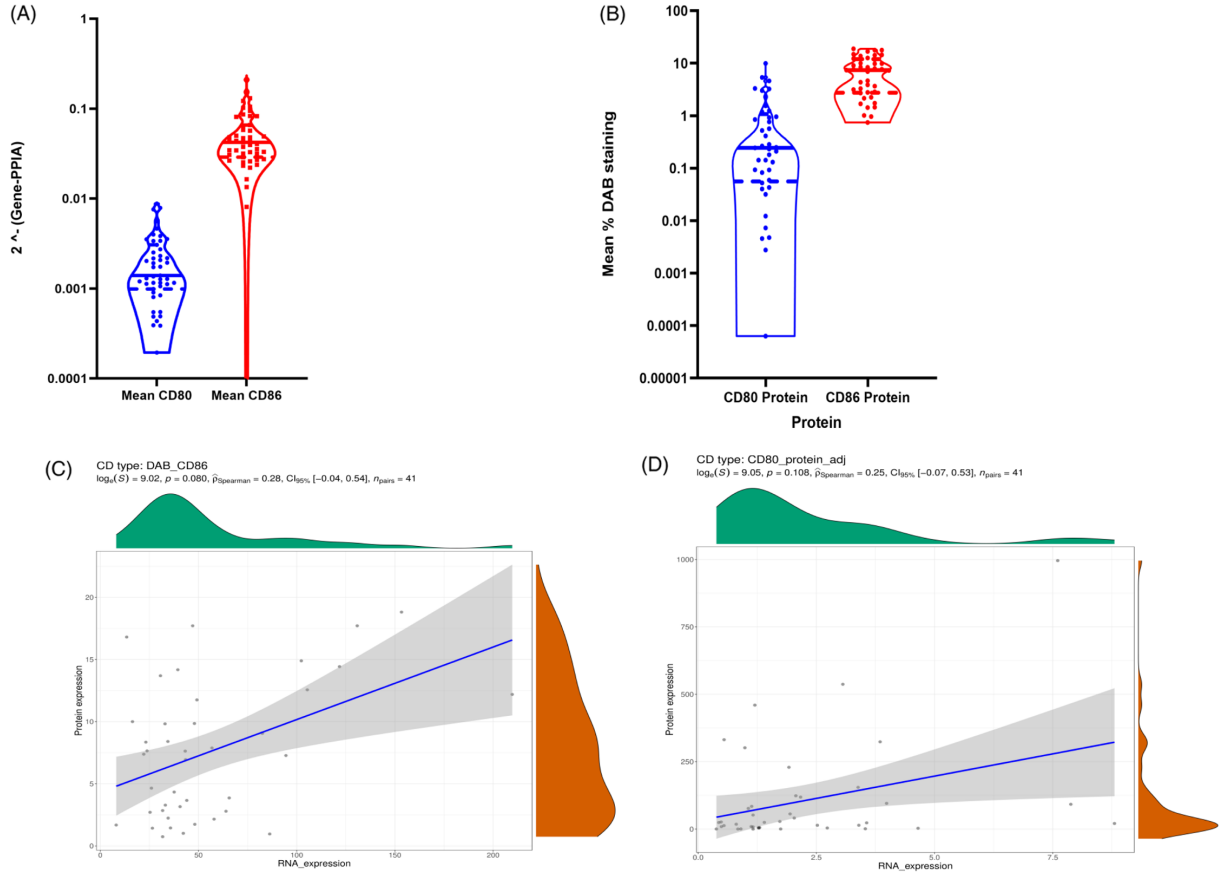


Figure 1: (A) Violin plot to visualize the data distribution of CD80 and CD86 mRNA expression in ONT database; (B) shows CD80 and CD86 protein expression in ONT database. (C–D) Spearman correlations between CD86 protein values and CD86 RNA values. (C) represents CD86 protein quantification based on the mean percentage of positive DAB signals correlation with mRNA values. (D) shows CD80 protein values quantified based on the mean percentage of positive DAB signals correlation with mRNA values.

Ipilimumab – anti-CTLA4 – has also shown responses in patients with brain metastases, highlighting efficacy within the central nervous system.[14] Expression of the most studied immune checkpoint proteins, programed death-ligand (PD-L1), was inversely correlated with OS in GBM patients.[15] However, the expression of CD80 and CD86 in GBM tissues and their prognostic significance in the tumor microenvironment (TME) of newly diagnosed GBM patients has not been reported yet. This study investigated the mRNA and protein expression of CD80 and CD86 in the TME of newly diagnosed GBM patients, aged below 70 years old and with KPS above 70% treated with the standard of care. In addition, this study highlighted a possible correlation between CD80 and CD86 expression and the immune cell populations in the TME of newly diagnosed GBM patients.

Materials and Methods

Patient Samples

OncoNeuroTek (ONT) is a local brain tumor tissue bank collecting samples from patients operated at the University Hospital La Pitié-Salpêtrière. All samples were collected with informed consent from patients. The inclusion criteria of the discovery local cohort (47 patients) were as follows: (i) newly diagnosed and histologically verified GBM, (ii) age at diagnosis is below 70 years, (iii) KPS above 70%, (v) known *MGMT* promoter methylation status, (vi) known IDH status, (vii) treated with the standard first-line therapeutic regimen including maximal safe surgery, chemoradiation and adjuvant temozolomide, and (viii) a documented clinical follow-up. The validation cohort (121 patients, TCGA cohort) clinical information and RNA-sequencing data (read counts) were downloaded from the National Cancer Institute's Genomic Data commons Data portal and from the NCBI GEO GSE62944, respectively. Similar inclusion criteria were used for both cohorts.

Immunohistochemistry Staining

Paraffin-embedded tissue blocks (5–7 µm) from biopsies of newly diagnosed GBM patients were received from the ONT biobank. The slides were obtained from diagnostic blocks and were selected to get a homogeneous group of patients for prognostic studies. Indeed, we have selected the patients aged below 70 years old, with a KPS > 70% and treated with the standard of care to be in line with inclusion criteria of the clinical trial that has established the

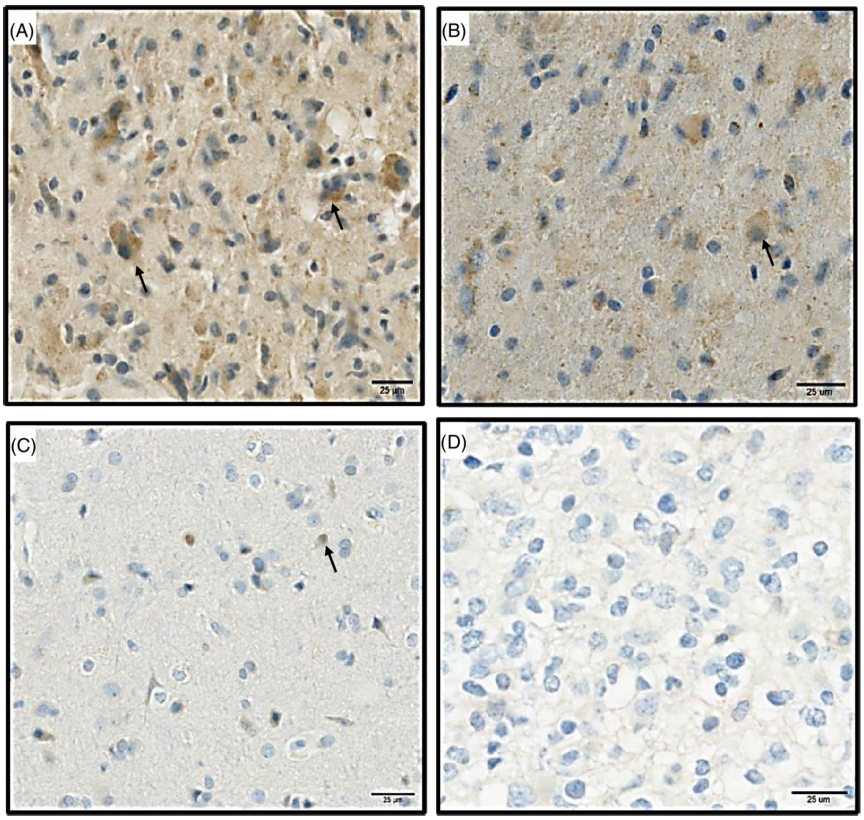


Figure 2: Represents the protein expression of CD86 and CD80 proteins in paraffin sectioned GBM samples. (A) High expression of CD86 protein. (B) Low expression of CD86. (C) High expression of CD80. (D) Low expression of CD80. Black arrows (brown signals) highlight a positive staining for CD80 and CD86 proteins and represent the signals that were used for quantifications, blue staining correspond to hematoxylin dye which was used as counterstaining.

standard of care.[4] Tissue sections (two sections per patients) were deparaffinized using xylene and rehydrated. For antigen retrieval, each slide was embedded in citrate buffer at pH 4.0 and heated for 15 min in the microwave at 800 W. 10% goat serum with 5% fetal bovine serum in 0.2% triton phosphate buffer saline was used as a blocking buffer. 3% hydrogen peroxide was used to block tissue peroxidation. Antihuman CD80 antibody (A16039; Abclonal) and antihuman CD86 antibody (A2353; Abclonal) were used at 1:500 dilution in blocking solution and incubated on the tissue slides overnight at room temperature. Avidin–Biotin Complex kit was used as a signal enhancer before the incubation in 3,3′-Diaminobenzidine (DAB). Slides were embedded in hematoxylin dye and rinsed with tap water for nuclear staining; gradual alcohol and xylene baths were used for dehydration and mounted with a hydrophobic mounting medium (Sigma, 24845633). All stained tissues were scanned via ZEISS Axio Scan 40× for bright field imaging.

Quantification of IHC Staining

Following all slides' imaging, three regions of interest with known dimensions (528 * 528 µm) were randomly selected for each tissue section and quantified using an in-house quantification Fiji code. Shortly, each image was imported to the Fiji program.[16] Using the color deconvolution tool, the area positive for DAB staining was isolated and quantified using a semiautomated in-house generated code. The percentage of DAB positive areas was calculated, and the mean value from the three images was calculated and used in the survival analysis.

Quantitative Reverse Transcriptase Polymerase Chain Reaction

RNA samples were obtained from ONT bank and used to synthesize cDNA. Reverse transcription of RNA samples was performed using the Maxima First Strand cDNA Synthesis Kit (Thermo Scientific, K1442) according to the manufacturer's recommendation with 100–250 ng of RNA. Quantitative reverse transcriptase polymerase chain reaction was used to quantify the expression levels of CD80 and CD86 in patients. PPIA gene was used as a house reference gene for normalization as previously described.[17] Primers were designed using Universal Probe Library (UPL) for Human. Primer's sequences were as follows: PPIA (left: atgctggacccaacacaaat; right: tctttcactttgccaaacacc; UPL probe 48) CD80 (left: gaagcaaggggctgaaaag; right: ggaa gttcccagaagaggtca; UPL probe 10) and CD86 (left: cagaagcagc-caaaatggat; right: gaatcttcagaggagcagcac; UPL probe 15). cDNA samples were analyzed using the Light Cycler Probe Master mix 2× (Roche, 04887301001) and the UPL detection system (Roche, 04483433001) in a Light Cycler 96 (Roche). For each qPCR, two independent experiments were completed with duplicate samples in each experiment. The mean of $2^{-(CT^{\text{gene of interest}}\_CT^{PPIA})}$ from the two different experiments was used in all analyses.

Statistical Analysis

A violin plot was used to visualize our data's full distribution (GraphPad Prism).[14] Spearman correlation between the expression values (RNA or protein) and age was evaluated to discard

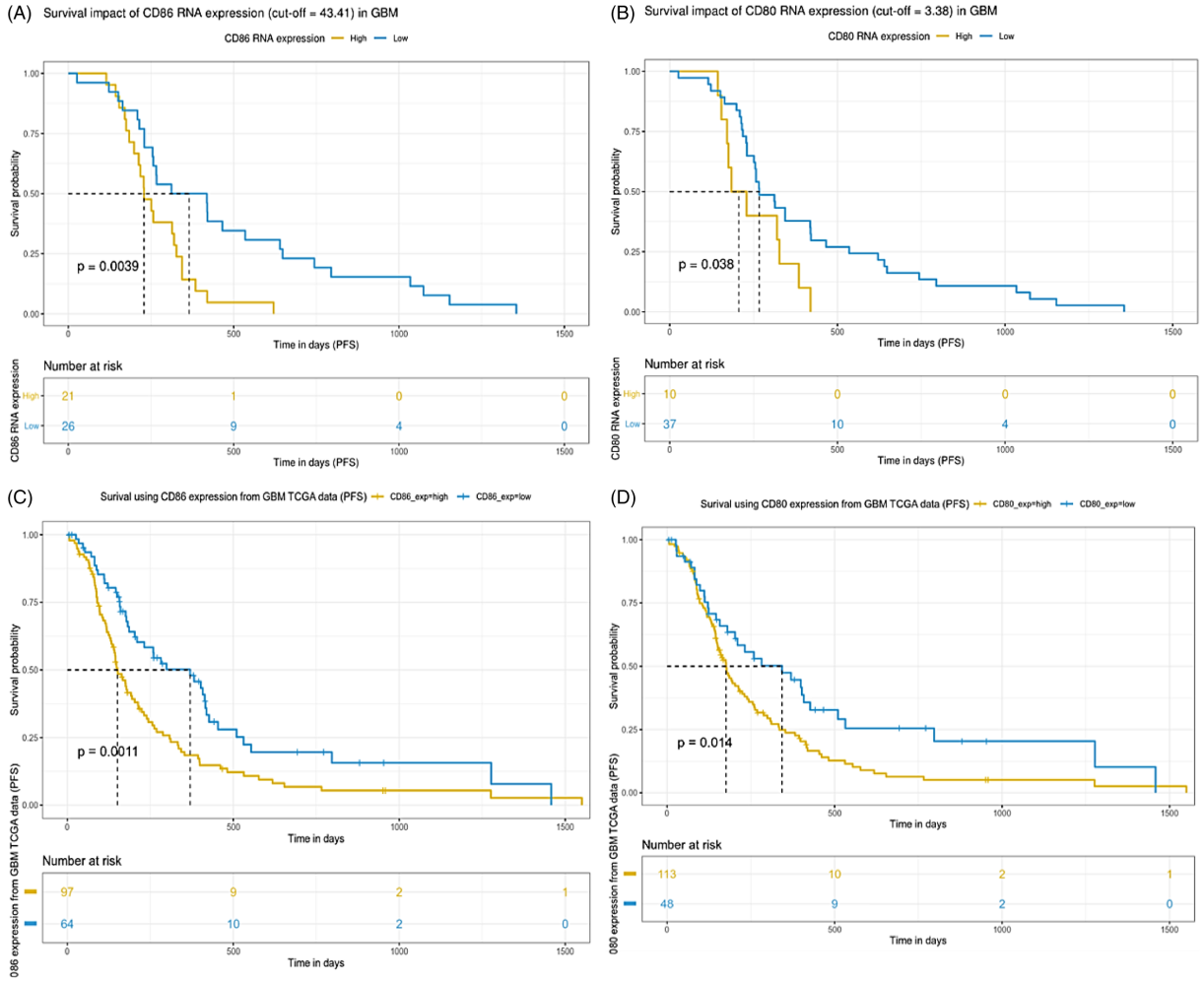


Figure 3: CD80 and CD86 mRNA expression and outcome in GBM in both ONT and TCCA database. (A) Kaplan–Meier PFS estimates in GBM patients in relation to CD86 (ONT database) (B) Kaplan–Meier PFS estimates in GBM patients in relation to CD80 (ONT database). (C) Kaplan–Meier PFS estimates in GBM patients in relation to CD86 (TCGA database). (D) Kaplan–Meier PFS estimates in GBM patients in relation to CD80 (TCGA database).

age bias. Survival analysis was performed by an open-source validated approach[18,19] by finding a supervised cutoff value for the CD80 or CD86 expression independently using the "survminer::surv_cutpoint" function, which determines the cut point based on the highest/lowest value of the log-rank statistics (low or high expression values), and then using these categories for Kaplan–Meier analysis or Cox proportional hazard regression modeling testing at each variable independently or to adjust for multiple variables including CD80/CD86 expressions and *MGMT* promoter methylation status *p*-values lower than 0.05 were considered significant.[20,21] Furthermore, we have used TCGA database to evaluate and profile tumor infiltrating immune populations and whether it differ among the highly expressed C86 tumor cells. TCGA immune data (i.e., CIBERSORT calculated immune populations) was retrieved from https://cavei.github.io/example-datasets/panCancerAnnotation.RData. Comparisons were performed by two-side Wilcoxon-test and *p*-values were corrected for multiple comparisons using FDR method.

Results

Patients and Tumors Characteristics

Forty-seven patients with a confirmed GBM diagnosis fulfilled the inclusion criteria: 14 men and 33 women (percentage 29.8%–70.2%). The patients' median age at diagnosis was 55.9 years (range: 24.3–69.5 years). KPS was 70 and above in all patients. The median OS is 559 days (range 31–2539), and the median PFS is 266 days (range 26–1355). The IDH status was evaluated as mutant for two patients (4.3%) while wildtype for 45 (95.7%). Furthermore, *the MGMT* promoter was methylated in 16 patients (34%) and unmethylated in 31 (66%). All patients were treated with the standard of care first-line treatment including maximal safe surgery, radio chemotherapy, and adjuvant chemotherapy with temozolomide.

CD80 and CD86 Expression at mRNA and Protein Level

At the mRNA level, CD86 expression was quantitatively higher than CD80 expression in the TME (Figure 1A). In agreement with

Table 1: Univariate analysis (Cox-P regression) for OS in both ONT and TCGA database

Characteristics		ONT					TCGA				
		N = 47	Percentage %	median OS (days)	*P*-value	HR [95% CI]	N = 121	Percentage %	median OS (days)	*P*-value	HR [95% CI]
MGMT	Methylated	16	34.04	986.5	**0.00032**	0.266 [0.129–0.547]	50	41.32	457	**0.0066**	0.544 [0.350–0.844]
	Unmethylated	31	65.95	441			71	58.67	273		
IDH	Wildtype	45	95.74	502	0.321	2.062 [0.493–8.623]	113	93.38	333	**0.0045**	5.39 [1.69–17.22]
	Mutant	2	4.25	1220			8	6.61	845		
CD80 mRNA	High	5	10.63	488	0.192	0.525 [0.200–1.382]	104	85.95	306	0.07	0.573 [0.314–1.046]
	Low	42	89.36	585			17	14.04	485		
CD86 mRNA	High	31	65.95	568	0.09	0.55 [0.27–1.11]	36	29.75	421	0.376	1.223 [0.783–1.911]
	Low	16	34.04	500			85	70.24	333		
		N = 41	Percentage %	median OS (days)	*P*-value	HR [95% CI]					
CD80 protein	High	8	19.51	950	**0.011**	3.53 [1.34–9.33]					
	Low	33	80.48	470							
CD86 protein	High	24	58.53	486	0.202	1.537 [0.794–2.972]					
	Low	17	41.46	568							

Table 2: Univariate analysis (Cox-P regression) for PFS in both ONT and TCGA database

Characteristics		ONT					TCGA				
		N = 47	Percentage %	Median PFS (Days)	*P*-value	HR [95% CI]	N = 121	%	Median PFS (Days)	*P*-value	HR [95% CI]
MGMT	Methylated	16	34	587.5	**0.00013**	5.12 [2.22–11.8]	50	41.32	194	**0.0095**	1.788 [1.15–2.77]
	Unmethylated	31	66	251			71	58.67	157		
IDH	Wildtype	45	95.7	266	0.407	0.54 [0.128–2.30]	113	93.38	158	**0.0117**	4.467 [1.40–14.3]
	Mutant	2	4.3	242.5			8	6.61	488		
CD80 mRNA	High	10	21.27	206.5	**0.0426**	0.464 [0.221–0.975]	80	66.11	156	**0.0428**	0.621 [0.392–0.985]
	Low	37	78.72	267			41	33.88	203		
CD86 mRNA	High	21	44.68	229	**0.0049**	0.38 [0.199–0.75]	72	59.50	145	**0.00283**	0.509 [0.327–0.793]
	Low	26	55.31	365.5			49	49	210		
		N = 41	Percentage %	Median PFS (Days)	*P*-value	HR [95% CI]					
CD80 Protein	High	25	60.97	229	0.0841	0.565 [0.296–1.08]					
	Low	16	39.02	402							
CD86 Protein	High	13	31.70	218	**0.0429**	0.48 [0.244–0.977]					
	Low	28	68.29	329							

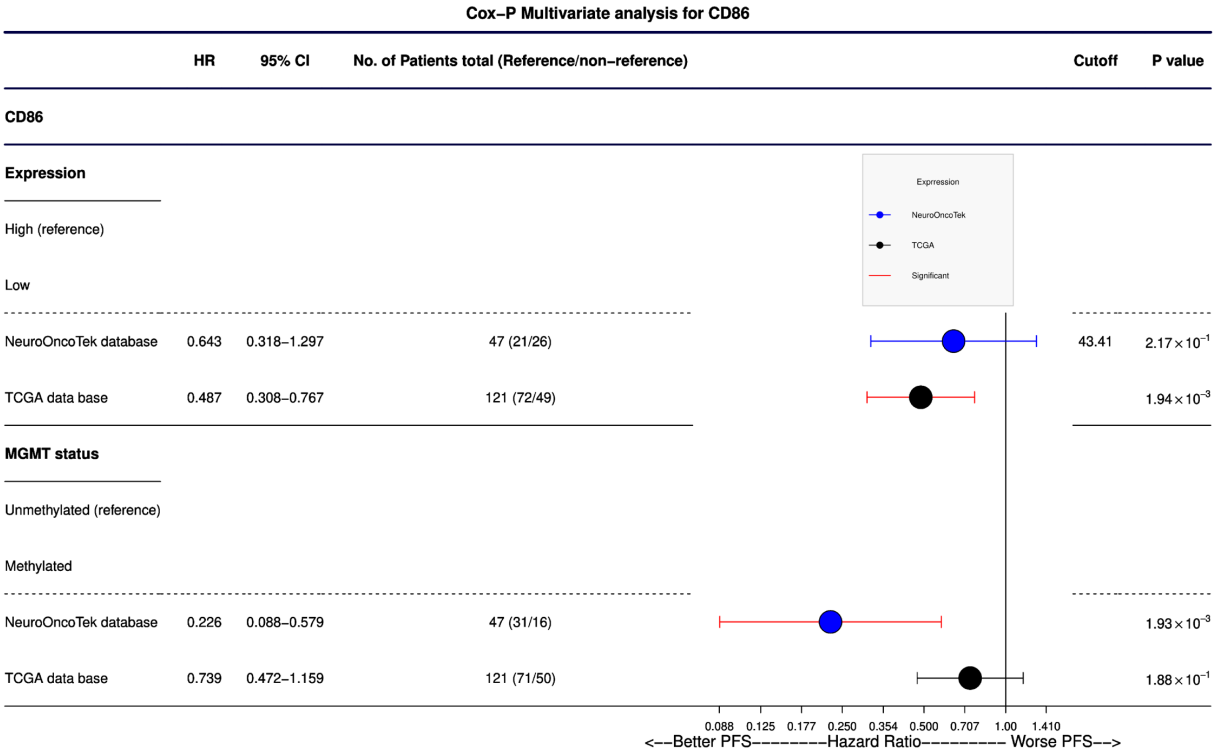


Figure 4: Cox-P (proportional hazards) multivariate analysis of CD86 protein expression and mRNA expression. CD86 was found to be an independent prognostic factor in TCGA database (p = 0.0019); mRNA expression of CD86 is a more predictive prognostic factor than MGMT methylation. A nonsignificant trend was observed in our ONT cohort.

mRNA expression, immunohistochemistry (IHC) analysis showed that the expression of CD86 is higher than CD80 in our discovery cohort (Figure 1B). Based on the IHC staining, CD80 and CD86 are observed in the cell membrane and/or the cytoplasm (Figure 2). Following protein quantification, we observed a positive correlation between RNA and protein expression for CD86 (Spearman coefficient of correlation Rho = 0.28; $p = 0.08$; Figure 1C). However, we observed a weaker correlation between mRNA and protein expression for CD80 ($p = 0.108$; Rho = 0.25; Figure 1D).

Prognostic Value of CD80 and CD86 Expression

Our patient's cohort was used as a discovery cohort, while the TCGA dataset was used as a validation cohort. In a univariate analysis, mRNA expression of CD80 and CD86 was not significantly correlated with OS in both the ONT cohort and TCGA dataset (Table 1). On the other hand, CD80 and CD86 mRNA high expression was significantly associated with shorter PFS ($p = 0.04$ and $p = 0.005$, respectively; Figure 3A,B). Moreover, these findings were validated using the TCGA cohort; higher CD80 and CD86 expressions were correlated with shorter PFS (*p*-value; 0.0428, 0.00283; Figure 3C,D). Interestingly, higher CD86 protein expression was associated with shorter PFS in the ONT cohort ($p < 0.005$; Table 2). CD80 and CD86 protein expression were not available in the TCGA dataset for validation purposes. However, we have used TCGA database to profile tumor-infiltrating immune cells in the selected cohort.

As expected, *MGMT* promoter methylation was associated with longer PFS and longer OS in the ONT cohort ($p < 0.05$ and $p < 0.05$, respectively) and TCGA dataset ($p < 0.05$ and $p < 0.05$, respectively) (Tables 1 and 2). Furthermore, *IDH* mutations were also associated with better OS and PFS in the TCGA database ($p < 0.05$ and $p < 0.05$, respectively); however, in the ONT cohort, the limited number of IDH-mutant GBM did not allow a robust analysis (n = 2). In multivariate analysis, CD80 mRNA expression did not provide additional prognostic information to *MGMT* promoter methylation in the ONT cohort. On the other hand, multivariate analysis of CD86 mRNA expression was an independent prognostic factor for PFS in the TCGA dataset only ($p < 0.05$; Figure 4). We have observed a similar trend ($p = 0.27$; Figure 4) in the ONT cohort, yet the trend was not significant, which could be related to the lower number of patients (n = 47) in the ONT cohort compared to (n = 121) in the TCGA database.

The Relationship Between CD86 Expression and Immune Cell Populations

Immune cell populations were evaluated using CIBERSORT, and we compared the immune cell populations between patients expressing both CD80 and CD86 as high and low expression. CD80 and CD86 are expressed on the surface of tumor-associated macrophages' surface suggesting a role in immunosuppressive TME. Immune cell population analysis showed low fraction of classically activated macrophages (M1) and higher fraction of immunosuppressive macrophages (M2). High CD86 expression group contained more patients with high M2 macrophages fraction ($p = 0.00013$; Figure 5). On the other hand, high CD86 expression group contained more patients with low tumor-infiltration

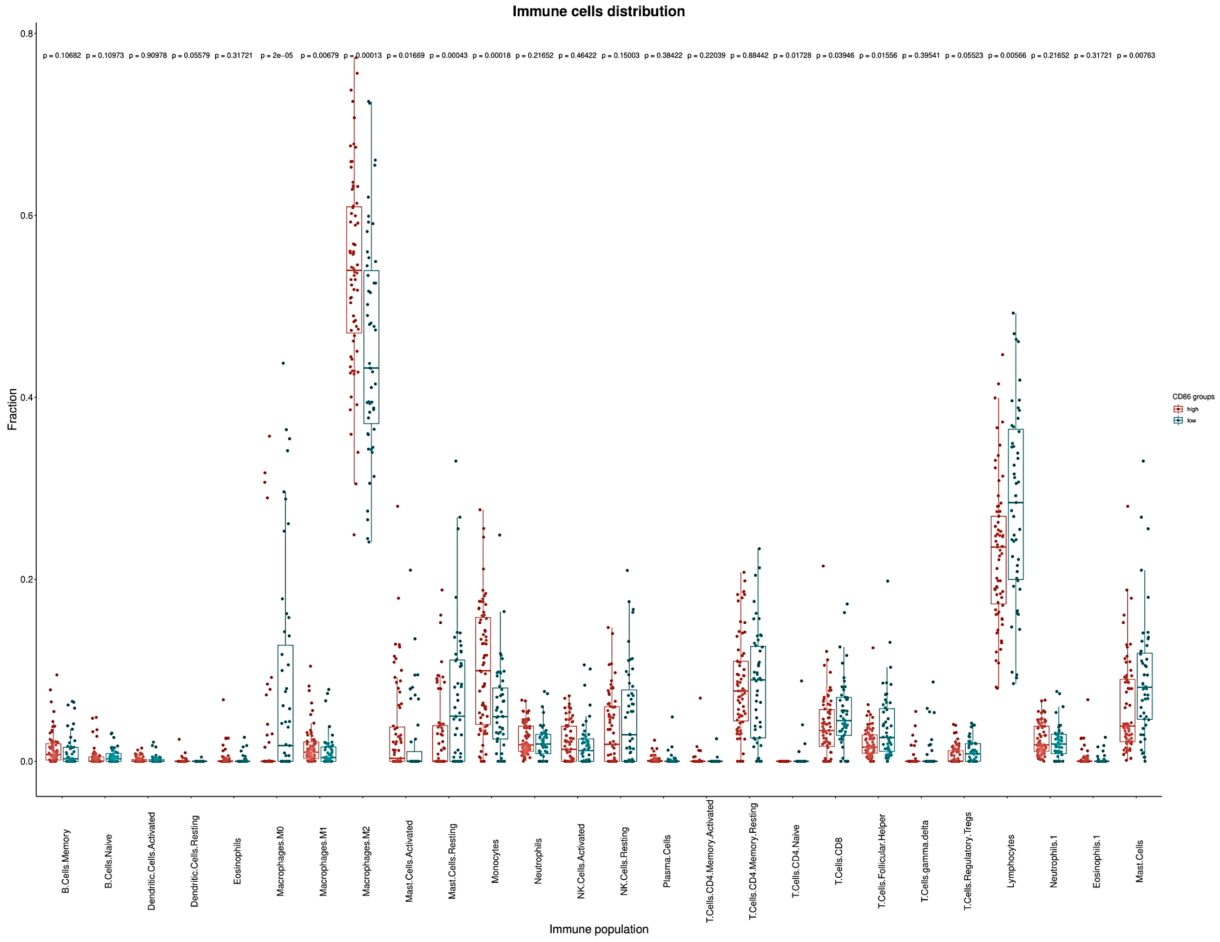


Figure 5: CIBERSORT calculated tumor infiltrating immune populations in TCGA database. Immune cell populations represented fraction of the X-axis immune cells to the whole gene expression mixture. Box plots depicting the estimated relative fractions of immune cell types by GBM category according to CD86 expression. The Y-axis here show the relative proportion which can range from 0 to 1. Relative fraction estimates the percentage of a given cell population in the total tumor infiltrate. In our analyses immunosuppressive M2 macrophages and lymphocytes were the most frequently observed immune phenotypes.

lymphocytes fraction ($p = 0.005$; Figure 5). This effect was not observed in CD80 expression patients. Additionally, high CD86 expression group contained more patients with low $CD8^+$ cell fraction ($p = 0.039$; Figure 5) whereas, the low CD86 expression group contained more patients with high $CD8^+$ fraction. Although further studies are warranted, these data suggest association between high CD86 expression, immunosuppressive TME, and low activity of $CD8^+$ cytotoxic T lymphocytes.

Discussion

CD80 and CD86 molecules play an essential role in influencing the immune recognition of GBM cells. They bind to the CD28 molecule with a costimulatory signal for T-lymphocytes activation. On the other hand, they bind to CTLA-4, resulting in an immunosuppressive effect. CTLA-4 has a higher affinity to CD80 and CD86, making these molecules' role in immunosuppressive effect higher than their costimulatory effect.[21] The current study has linked CD80 and CD86 expression on GBM TME to PFS. We observed a low correlation between mRNA and protein expression of CD80. However, a better correlation was observed between CD86 protein and mRNA expression. Low correlation between the mRNA and protein expression might be due to posttranscriptional mechanisms involved in turning mRNA into protein. Not to mention, there is a possible error and noise in protein quantification and mRNA extraction that could influence mRNA stability and protein expression.[20] In addition to DAB staining intensity used in our study, quantification of protein using the number of positive cells should also be evaluated in future IHC analyses to better understand expression of proteins and mRNA of interest.

Number of patients (n = 47) in the ONT cohort is lower than the number of patients in the TCGA dataset (n = 121). The higher number of TCGA GBM samples could be one reason that affected the statistical analysis and provided a better prognostic value than the ONT cohort. Indeed, GBM samples' availability with comprehensive clinical and biological annotations and fulfilling the inclusion criteria is a limitation for a larger cohort. Larger patient cohort is needed to evaluate the prognostic value of CD86 expression in the TME of GBM patients. Using TCGA data to profile immune cell populations interestingly revealed that CD86 expression is associated with an immunosuppressive TME with low activity of cytotoxic T cells however protein analysis of immune cell

populations is needed to validate our findings from TCGA immune cell population profiling. Indeed, high CD86 expression is associated with a cold immune microenvironment with a limited antitumor immune response promoting tumor growth and poor prognosis.

The expression of 50 immune checkpoint molecules was investigated in breast cancer. The study showed that high expression of costimulatory immune checkpoint molecules was associated with better PFS. However, no significant effect on prognosis was associated with CD80 and CD86 expression in the selected cohort.[22] Feng et al.[23] reported that low expression of CD80 is a predictive biomarker for poor prognosis in gastric adenocarcinoma. Furthermore, CD80 and CD86 were found to be potential biomarkers for better prognosis survival in nasopharyngeal carcinoma.[24] Additionally, the molecular characterization of PDL1 expression was correlated with other checkpoint proteins, that is, CD80, highlighting that higher levels of immunosuppression are associated with GBM than lower-grade gliomas (LGG).[25] In myeloma cell lines, silencing the CD28–CD86 pathway resulted in myeloma cells' significant cell death.[26] A recent study constructed a more robust model, using GBM and LGG data from the TCGA and Chinese Glioma Genomic Atlas, and identified that low expression of CD86 molecules is a good prognostic indicator for OS. PFS analysis was not applied in this study.[27]

In 2017, Berghoff et al. described a specific signature to predict the success of TMZ in *MGMT*-methylated patients. They showed that the TME signature could be used to indicate an individual's TMZ sensitivity. The TME was identified to be different between *IDH*-mutant and *IDH*-wildtype. A richer tumor infiltrative lymphocyte and a higher expression of PDL1 were observed in *IDH*-wildtype tumors.[28] However, to date, no studies have linked *MGMT* promoter methylation with the TME. A recent research article has studied the expression of immune checkpoint inhibitor Tim3 and *MGMT* methylated status. They identified that a high expression of Tim3 in *MGMT*-unmethylated patients is linked to poor prognosis.[29] Pratt et al.[30] have reported that the expression of PD-L1 is a negative prognostic biomarker in recurrent *IDH*-wildtype GBM. In line with these findings, our study supports that the expression of immune checkpoint inhibitors may inhibit T-lymphocyte and antitumor reaction. A recent integrated analysis of the prognostic value of CD86 reveals that CD86 is heterogeneously expressed in gliomas and is an independent unfavorable prognostic value in LGG.[31]

CD86 molecular status could be explored as a predictor of response to immunotherapies in the setting of future clinical trials dedicated to GBM patients. Our study suffers from the limitation of retrospective studies with a limited number of patients. Nonetheless, our results were validated in an independent dataset and support investigations of immune checkpoint molecules as potential prognostic biomarkers and potential predictive biomarkers of response to immunotherapies in GBM.

Acknowledgements. This work was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement #766069 (GLIO-TRAIN).

Conflicts of Interest. No potential conflicts of interest were disclosed.

Author Contributions. MA, AI designed the experiments, wrote the manuscript, and approved the manuscript's final version. MA performed the experiments. IHV performed the statistical analysis and revised the manuscript. FB, JL, MV provided a technical support for IHC optimization and protein quantification. All authors reviewed the manuscript.

Ethics Approval. All samples were collected with informed consent from patients.

Data Statement. All annotated data will be available upon request from the authors.

References

1. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016;131:803–20.
2. deSouza RM, Shaweis H, Han C, et al. Has the survival of patients with glioblastoma changed over the years? Br J Cancer. 2016;114:146–50.
3. Marenco-Hillembrand L, Wijesekera O, Suarez-Meade P, et al. Trends in glioblastoma: outcomes over time and type of intervention: a systematic evidence based analysis. J Neuro-Oncol. 2020;147:297–307.
4. Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med. 2005;352:987–96.
5. Leven C, Padelli M, Carré JL, Bellissant E, Misery L. Immune checkpoint inhibitors in melanoma: a review of pharmacokinetics and exposure-response relationships. Clin Pharmacokinet. 2019;58:1393–405.
6. Muftuoglu Y, Liau LM. Results from the CheckMate 143 clinical trial: stalemate or new game strategy for glioblastoma immunotherapy? JAMA Oncol. 2020;6:987–89.
7. Ville S, Poirier N, Blancho G, Vanhove B. Co-stimulatory blockade of the CD28/CD80-86/CTLA-4 balance in transplantation: impact on memory T cells? Front Immunol. 2015;6:411.
8. Trombetta AC, Soldano S, Contini P, et al. A circulating cell population showing both M1 and M2 monocyte/macrophage surface markers characterizes systemic sclerosis patients with lung involvement. Respir Res. 2018;19:186.
9. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347:1260419.
10. Hersey P, Si Z, Smith MJ, Thomas WD. Expression of the co-stimulatory molecule B7 on melanoma cells. Int J Cancer. 1994;58:527–32.
11. Wei SC, Duffy CR, Allison JP. Fundamental mechanisms of immune checkpoint blockade therapy. Cancer Discov. 2018;8:1069–86.
12. Rowshanravan B, Halliday N, Sansom DM. CTLA-4: a moving target in immunotherapy. Blood. 2018;131:58–67.
13. Letendre P, Monga V, Milhem M, Zakharia Y. Ipilimumab: from preclinical development to future clinical perspectives in melanoma. Future Oncol. 2017;13:625–36.
14. Savoia P, Astrua C, Fava P. Ipilimumab (Anti-Ctla-4 Mab) in the treatment of metastatic melanoma: effectiveness and toxicity management. Hum Vaccin Immunother. 2016;12:1092–101.
15. Nduom EK, Wei J, Yaghi NK, et al. PD-L1 expression and prognostic impact in glioblastoma. Neuro-Oncol. 2016;18:195–205.
16. Schindelin J, Arganda-Carreras I, Frise E, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012;9:676–82.
17. Vessières A, Quissac E, Lemaire N, et al. Heterogeneity of Response to Iron-Based Metallodrugs in Glioblastoma Is Associated with Differences in Chemical Structures and Driven by FAS Expression Dynamics and Transcriptomic Subtypes. Int J Mol Sci. 2021;22(19). .
18. Zhou R, Zeng D, Zhang J, et al. A robust panel based on tumour microenvironment genes for prognostic prediction and tailoring therapies in stage I-III colon cancer. EBioMedicine. 2019;42:420–30.
19. Li S, Chen S, Wang B, Zhang L, Su Y, Zhang X. A robust 6-lncRNA prognostic signature for predicting the prognosis of patients with colorectal cancer metastasis. Front Med. 2020;7:56.
20. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. 2003;4:117.
21. van Nieuwenhuijze A, Liston A. The molecular control of regulatory T cell induction. Prog Mol Biol Transl Sci. 2015;136:69–97.
22. Fang J, Chen F, Liu D, Gu F, Chen Z, Wang Y. Prognostic value of immune checkpoint molecules in breast cancer. Biosci Rep. 2020;40:BSR20201054.

23. Feng XY, Lu L, Wang KF, et al. Low expression of CD80 predicts for poor prognosis in patients with gastric adenocarcinoma. Future Oncol. 2019;15:473–83.
24. Chang CS, Chang JH, Hsu NC, Lin HY, Chung CY. Expression of CD80 and CD86 costimulatory molecules are potential markers for better survival in nasopharyngeal carcinoma. BMC Cancer. 2007;7:88.
25. Wang Z, Zhang C, Liu X, et al. Molecular and clinical characterization of PD-L1 expression at transcriptional level via 976 samples of brain glioma. Oncoimmunology. 2016;5:e1196310.
26. Gavile CM, Barwick BG, Newman S, et al. CD86 regulates myeloma cell survival. Blood Adv. 2017;1:2307–19.
27. Qiu H, Li Y, Cheng S, Li J, He C, Li J. A prognostic microenvironment-related immune signature via ESTIMATE (PROMISE Model) predicts overall survival of patients with glioma. Front Oncol. 2020;10:580263.
28. Berghoff AS, Kiesel B, Widhalm G, et al. Correlation of immune phenotype with IDH mutation in diffuse glioma. Neuro-Oncol. 2017;19:1460–68.
29. Zhang J, Sai K, Wang XI, et al. Tim-3 expression and MGMT methylation status association with survival in glioblastoma. Front Pharmacol. 2020;11:584652.
30. Pratt D, Dominah G, Lobel G, et al. Programmed death ligand 1 is a negative prognostic marker in recurrent isocitrate dehydrogenase-wildtype glioblastoma. Neurosurgery. 2019;85:280–89.
31. Qiu H, Tian W, He Y, et al. Integrated analysis reveals prognostic value and immune correlates of CD86 expression in lower grade glioma. Front Oncol. 2021;11:654350.

Synthèse en français

Introduction

Le lymphome primitif du système nerveux central (LPSNC) est un sous-type rare de lymphome diffus à grandes cellules B (LDGCB) situé dans le SNC avec un pronostic moins favorable. Alors que le LDGCB constitue 25 à 35% des lymphomes non hodgkiniens (LNH) chez l'adulte, on estime que le PCNSL représente jusqu'à 1 à 2% des LNH, 4 à 6% de tous les lymphomes extranodaux et environ 2 à 3% de toutes les tumeurs du SNC. Bien qu'il ait été considéré pendant longtemps comme un LDGCB dans le SNC, principalement parce qu'il présente l'histologie d'un LDGCB, il a été prouvé que le LPSNC est une entité biologique différente sur le plan moléculaire. Comme le LPSNC dans la population immunocompétente représente la grande majorité des patients, avec une lymphomagenèse distincte de celle liée au LPSNC immunodéficient (impliquant le virus d'Epstein-Barr, $< 10\%$ des cas), ma thèse se concentre sur le LPSNC chez les patients immunocompétents. De plus, le pronostic du LPSNC est significativement plus mauvais que celui du LDGCB, avec une médiane de survie globale (SG) et de survie à 5 ans de 26 mois et 22% pour le LPSNC contre 124 mois et 51% pour le LDGCB systémique. Cependant, il existe une grande hétérogénéité au sein du LNPC et les raisons physiopathologiques sous-jacentes du comportement clinique des tumeurs ne sont pas encore élucidées.

Le diagnostic de référence du LPSNC, après les symptômes initiaux évocateurs et l'imagerie par résonance magnétique, est soit une biopsie stéréotaxique avec un examen pathologique, soit un examen du liquide céphalorachidien (LCR) (en cas d'atteinte leptoméningée), soit un examen du liquide vitré (en cas d'atteinte oculaire). L'immunohistochimie (IHC) aide l'histologie à poser le diagnostic de LPSNC, où presque toutes les cellules du LPSNC expriment des marqueurs pan-cellulaires B comme CD20, CD19, CD22, CD79A, IgM et IgD mais pas IgG. Les marqueurs de différenciation des cellules B, BCL-6 (60-80%) pour les cellules B du centre germinal (CG), et IRF4/MUM1 (90%) pour les cellules B

du CG tardif et les plasmocytes, sont également importants pour le diagnostic du LPSNC. Le LPSNC a donc été défini immunophénotypiquement comme des cellules B post-CG. De plus, les cellules ont une activité prolifératives très élevée puisque l'indice de prolifération Ki-67 est généralement >70% et peut même être >90%. Il est surprenant de constater que, même si la perte du complexe majeur d'histocompatibilité (CMH) se produit dans environ 50% des cas de LNPC, les HLA-A, HLA-B, HLAC et HLA-DR sont exprimés de manière variable.

Sur le plan pathologique, les cellules du LPSNC présentent une infiltration périvasculaire fréquente et consistent, sur le plan cytomorphologique, des cellules atypiques avec des noyaux ronds, ovales, irréguliers ou pléomorphes de taille moyenne à grande et des nucléoles distincts, correspondant à des centroblastes ou des immunoblastes. Alors que l'OMS qualifiait le LPSNC de "cellule B tardive de sortie du centre germinal arrêtée dans la différenciation terminale des cellules B qui partage des caractéristiques génétiques à la fois avec les cellules B activées (CBA) et les cellules B du centre germinal (BCG)," des études récentes sur le LPSNC combinant des altérations génétiques et des données d'expression transcriptomique ont révélé que les cellules du LPSNC étaient arrêtées à différents stades du transit du CG (de la zone sombre à la zone claire). En conséquence de cette permanence du CG, les cellules LPSNC présentent, par rapport à d'autres malignités à cellules B, une hypermutation somatique continue plus élevée qui conduit à une hypermutation somatique aberrante due à l'activité hors cible de l'enzyme cytidine désaminase induite par l'activation (AID, codée par le gène *AICDA*). En outre, les translocations du gène BCL2/BCL6 avec les régions régulatrices de l'IGH entraînent une surexpression ectopique de la protéine, provoquant finalement des réentrées itératives de CG et donc des expansions clonales, une instabilité génomique, l'acquisition de mutations (hors cible de l'AID) et la lymphomagenèse.

Le LPSNC se caractérise par une activité constitutive du facteur nucléaire-$\kappa\beta$ (NF-$\kappa\beta$), qui entraîne une prolifération cellulaire et une prévention de l'apoptose cellulaire, induite par des altérations des gènes de la voie BCR (*CD79B* dans 43%, *SHIP*-25%, *CBL*-4% et *BLNK*-4%), de la voie des récepteurs Toll-like (TLR) ($MYD88^{L265P}$-64%) et d'autres (*CARD11*-22%, *BCL2*-43%, *MALT1*-43%). Le complexe BCR (récepteur des cellules B), composé des chaînes lourdes/légères de l'IG (immunoglobuline) ainsi que des sous-unités CD79A et CD79B, est indispensable à la survie des cellules B puisqu'il induit la différenciation, la prolifération et l'apoptose de ces dernières. De plus, alors que la voie BCR transmet également ses signaux au complexe signalosome CARD11-BCL10-MALT1 (CBM), la bruton tyrosine kinase (BTK) relie les voies de signalisation BCR et TLR à l'activation en aval de NF-$\kappa\beta$. En outre, en 2018, les LPSNCs ont

été rapprochés des LDGCB dits “MCD” ou “Cluster 5” (C5), tous deux convergeant sur la cooccurrence de mutations *MYD88*^{L265P} et *CD79B*, et la présence de pertes de copies impliquant 6p21-22 (échappement immunitaire médié par une perte *HLA*), 6q21 et 9p21.3 (perte bialélique *CDKN2A*). Les conséquences transcriptomiques de ces altérations au sein du sous-type MCD ont été décrites ultérieurement par Wright et al. (2020), qui ont constaté une augmentation de la prolifération (induite par la perte bialélique de *CDKN2A*), de l'activité NF-$\kappa\beta$ (induite par les mutations *MYD88*^{L265P} et *CD79B*), de l'activité de la kinase JAK1, de l'expression des IgM et de la signalisation BCR active chronique dépendant de l'auto-antigène (forte expression des $\text{IgV}_{\text{H4-34}}$).

Le microenvironnement tumoral (MET) joue un rôle inéluctable et remarquable dans la biologie des tumeurs et se définit comme un réseau cellulaire (c'est-à-dire les vaisseaux sanguins, les cellules immunitaires et les fibroblastes), moléculaire (c'est-à-dire les molécules de signalisation intercellulaire, la matrice extracellulaire) et dynamique entourant les cellules tumorales. De plus en plus de preuves suggèrent que le développement tumoral n'est pas seulement dû à l'accumulation d'anomalies intrinsèques mais aussi à des signaux extrinsèques provenant du MET. Dans le contexte du LDGCB, une première étude de 2021 a révélé l'existence de quatre MET différents présentant des comportements cliniques distincts. Les sous-types découverts, appelés GC-like, mesenchymal, inflammatory et depleted, présentent de nombreuses caractéristiques tumorales chaudes et froides et partagent certaines altérations génétiques inter-tumorales par groupe. Le sous-type GC-like se caractérise par la présence de cellules endothéliales lymphatiques (CEL), de cellules dendritiques folliculaires, de cellules T auxiliaires folliculaires (TFH), et de lymphocytes T régulateurs (Tregs) dans le MET, ainsi que par des altérations génétiques de *TNFRSF1*, *CD83*, *STAT6*, et *HSF1*. Le sous-type mésenchymateux présentait une présence plus importante de cellules endothéliales vasculaires, de fibroblastes associés au cancer, de cellules réticulaires des fibroblastes (FRC) et de macrophages M1 (pro-inflammatoires) dans le MET; en outre, il présentait également des mutations dans *E2H2*, *B2M*, *GNA13*, *GNAI2* et *P2RY8*. Le sous-type déplété présentait des altérations génomiques entraînant une diminution de l'activité de p53, une perturbation de la régulation du cycle cellulaire (par exemple, des délétions de *CDKN2A*) et une activité prolifératives élevée. Enfin, la tumeur de type chaud (sous-type inflammatoire) était enrichie en neutrophiles, en macrophages associés aux tumeurs, en macrophages M1, en Tregs, en TFHs, en cellules T CD8^+ avec une forte expression de PD-1 (épuisé), et également en activités cellules tueuses naturelles (NK), MHC-I, molecules de point de contrôle immunitaire (IC), NF-$\kappa\beta$, JAK/STAT, et TNF.

Dans le contexte du LPSNC, la localisation précise de la tumeur détermine probablement la composition du MET puisqu'elle peut se développer dans le parenchyme cérébral, mais aussi dans les espaces périvasculaires et méningés. Le LPSNC est considéré comme une tumeur confinée dans une zone "immuno-privilégiée" en raison de la présence de la barrière hémato-encéphalique. Néanmoins, des études récentes ont décrit un réseau de vaisseaux lymphatiques parallèles aux sinus veineux duraux qui permettent le drainage des cellules et du LCR vers les ganglions cervicaux profonds. Ces vaisseaux lymphatiques du SNC expriment toutes les caractéristiques moléculaires des CELs et transportent environ 24% de toutes les cellules T sinusales et 12% de toutes les cellules CMH-II$^+$ sinusales. Alors que les premières études ont établi un lien entre l'amélioration de la survie et les cellules T périvasculaires ou les lymphocytes infiltrant la tumeur (TILs en anglais), des études récentes (utilisant des données RNA-seq) ont découvert que ces TILs exprimaient des molecules de contrôle immunitaire (IC) comme PD-1 et TIM-3. Il est intéressant de noter que, même si des études antérieures ont trouvé des amplifications récurrentes du gène 9p24.1 (impliquant PD-L1), Marcelis et al. n'ont pas trouvé de telles amplifications au moyen de la méthodologie FISH. Par ailleurs, une augmentation globale du ratio de TAMs de type M1/M2 a été associée à un meilleur résultat en utilisant l'IHC ou la déconvolution immunitaire assistée par ARN. D'autre part, une analyse transcriptomique récente combinant l'ARN-seq (n = 20) et les microréseaux (n = 34) a décrit trois immunophénotypes appelés riche, intermédiaire et pauvre avec des implications sur le système d'exploitation. Alors que le groupe riche en immunité présentait la meilleure SG et était caractérisé par un nombre élevé de CD4$^+$, CD8$^+$, Tregs, TAMs et cellules dendritiques (DCs), le groupe pauvre en immunité était pratiquement une tumeur de type froid, et le groupe intermédiaire présentait une hétérogénéité des cellules immunitaires.

En ce qui concerne le traitement, le pilier pour les patients atteints de LPSNC nouvellement diagnostiqué est, selon les directives du National Comprehensive Cancer Network (NCCN) (2020), une chimiothérapie d'induction au méthotrexate à haute dose (HD-MTX) (1-8 g/m^2) suivie d'une radiothérapie du cerveau entier comme thérapie de consolidation. Cependant, la neurotoxicité à long terme dérivée de la chimiothérapie a conduit au développement de la polychimiothérapie, comme le HD-MTX, le rituximab, la vincristine et la procarbazine; le HD-MTX et le témozolomide; le HD-MTX avec autogreffe de cellules souches (ASCT). Bien que le LPSNC soit chimiosensible, 33% des patients sont réfractaires au traitement de première ligne, et jusqu'à 60% des patients finissent par rechuter 2 à 5 ans après le diagnostic initial.

Malgré les altérations moléculaires et les perspectives de traitement que ces études ont apportées, l'hétérogénéité de la réponse biologique et thérapeutique du LP-SNC n'a pas été correctement prise en compte, principalement en raison de l'absence d'un grand nombre de patients et de l'intégration de données multi-omiques, c'est-à-dire l'existence de types distincts d'informations moléculaires (par exemple, méthylation, mutations, alterations du nombre de copies, expression génique, localisation de la tumeur, MET, etc.) pour la même cohorte. Par conséquent, l'identification de sous-groupes de patients atteints de LPSNC ayant des facteurs biologiques communs de la maladie et de son issue clinique est d'une extrême importance pour adapter les stratégies de traitement. Cependant, l'identification de tels sous-groupes moléculaires est extrêmement difficile, principalement en raison de la grande hétérogénéité génétique, phénotypique et MET. Par conséquent, il existe un besoin non satisfait de réaliser de grandes études multi-omiques en vue de personnaliser les soins cliniques et d'améliorer les résultats des patients.

Objectifs

Cette thèse vise à caractériser le paysage multi-omique du LPSNC, y compris la génomique, l'épigénomique, la transcriptomique et la clinicomique, et à intégrer ces données pour trouver des sous-groupes moléculaires de LPSNC ayant une pertinence biologique et clinique. Par conséquent, les objectifs sont divisés en trois parties:

- **Chapitre 4.1:** Faire une revue de la littérature pour comprendre la structure/diversité HLA et la susceptibilité génétique dans le LPSNC et les autres LNH à cellules B.

- **Chapitre 4.2:** Développer un code pour suivre les mutations de c-AID et explorer leurs implications au niveau pan-cancer (~ 50 000 échantillons).

- **Chapitre 4.3:** Extraire, analyser et intégrer des données multi-omiques afin de trouver et de caractériser des sous-groupes moléculaires de LPSNC présentant des facteurs biologiques causaux communs de la maladie et de l'issue clinique.

Résultats

Structure/diversité HLA et susceptibilité génétique dans le LPSNC et autres LNH à cellules B

Les associations de risque des LNH à cellules B (y compris le LPSNC) ont été initialement attribuées aux antécédents familiaux de la maladie, à l'inflammation et aux composants immunitaires, y compris les variations génétiques HLA. Cependant, de récentes études d'association pangénomique ont permis d'obtenir davantage d'informations sur le sujet. Ici, je passe en revue la structure HLA et sa diversité et résume tous les articles originaux montrant des preuves de variations génétiques sur cinq sous-types de NHL.

Dans l'article de revue de la littérature, nous avons montré que les variants HLA sont les plus étudiés dans le contexte du LNH à cellules B puisque cette région est critique pour les réponses immunitaires innées et adaptatives. Il est intéressant de noter que le statut HLA s'est avéré être un facteur de risque dans les LNH à cellules B en favorisant l'échappement immunitaire, ce qui a également été observé spécifiquement pour le LPSNC. La production d'antigènes/néoantigènes est un important mécanisme d'échappement de la tumeur à la surveillance immunitaire, qui peut être perturbé par l'homozygotie HLA, comme c'est le cas dans la plupart des lymphomes étudiés (y compris le LPSNC).

De plus, en ce qui concerne spécifiquement le LPSNC, la seule étude évaluant les associations entre les variants génétiques et le risque de LPSNC a été réalisée par notre groupe dans une cohorte française. Bien que cette étude ait trouvé quelques variantes supplémentaires associées au risque de LPSNC, il est clair que des études supplémentaires sont nécessaires pour mieux élucider la pathogenèse du LPSNC.

Implications des mutations liées à l'AID au niveau pan-cancer

Comme nous l'avons vu tout au long de cette thèse, l'activité hors cible de AID s'inscrit dans le contexte de la biologie des cellules B et de la lymphomagenèse. Dans cette étude, qui est la plus importante à ce jour, j'ai intégré plus de 50 000 échantillons en vrac et 2,5 millions de cellules à résolution unicellulaire dans 80 types de tumeurs (y compris les tumeurs malignes à cellules B) et à différents niveaux de données. L'objectif principal, dans l'article, était de décrire en détail les implications oncogéniques et cliniques des mutations hors cible de c-AID à

l'échelle pan-cancer ; cependant, l'objectif principal de cette section concernant cette thèse était de développer et de valider le code pour cibler les mutations c-AID.

Tout d'abord, nous avons démontré que l'expression de *AICDA* n'est présente que dans les cellules B normales en utilisant une série d'études RNA-seq unicellulaires; néanmoins, cela change après une transformation maligne puisque nous avons observé son expression dans différents types de cancer à une résolution unicellulaire. Ensuite, nous avons évalué notre code de suivi des mutations c-AID à l'aide de motifs tétranucléotidiques en l'appliquant tout d'abord à une série de cancers hématologiques et en trouvant des cibles AID déjà signalées (par exemple, *PIM1*, *HIST1H1C*), en évaluant ensuite que notre code n'identifie pas les mêmes mutations que d'autres signatures somatiques COSMIC, et enfin, en excluant que les mutations AID observées soient générées par hasard. De plus, nous avons également décrit, comme prévu, que la fréquence des mutations c-AID est plus élevée dans les cancers hématologiques par rapport aux autres.

Après avoir validé le code, nous avons décrit le paysage et les implications des mutations de c-AID. Nous avons constaté que l'activité de c-AID se produit principalement pendant la transcription de ses gènes hors cible et qu'elle est accrue dans les tumeurs MSI. De plus, nous avons montré que dans certains types de cancer, l'activité promiscuous de c-AID vise les hotspots de sélection les moins positifs qui entrent en synergie avec des mutations hotspot plus fortes antérieures (mutation mineure PIK3CA E726, particulièrement présente dans le carcinome de la peau et le cancer du sein). Enfin, nous avons démontré que la fraction de mutations liées au AID est une valeur pronostique indépendante de l'ICI (immune chekpoint inhibition en anglais) et avons présenté différentes analyses pour expliquer ces résultats.

La recompilation de tous les ensembles de données publics, ainsi que les résultats et le code pour détecter les mutations c-AID, fournissent la base pour tester le rôle potentiel des mutations c-AID dans les cancers hématologiques et non hématologiques. Cependant, en raison de la nature bioinformatique de l'étude, plusieurs validations biologiques des résultats doivent être effectuées.

L'intégration de données multi-omiques révèle des sous-types moléculaires de LPSNC ayant une pathogenèse commune et des implications en termes de résultats cliniques

A notre connaissance, l'étude présentée dans le Chapitre 5.3 représente la plus grande étude multi-omique du LPSNC menée à ce jour. Notre étude

s'appuie sur la classification actuelle des LDGCB, les LDGCB MCD/C5, en ajoutant l'hétérogénéité moléculaire au sein du LPSNC qui peut informer sur sa pathogenèse et finalement donner des cibles thérapeutiques potentielles. Ici, j'ai trouvé quatre sous-types de LPSNC avec des caractéristiques multi-omiques partagées telles que des voies oncogéniques distinctes, des phénotypes d'expression génétique, des profils de méthylation, des MET et des caractéristiques clinico-radiologiques.

Nos résultats aident à élucider la réponse hautement hétérogène dans le LPSNC en connectant différentes couches multigénomiques avec les informations clinico-omiques. Nous avons montré ici que le groupe CS4 partage avec le groupe CS3 une activation constitutive de NF-$\kappa\beta$, qui est l'une des principales caractéristiques des sous-types MCD ou C5 du DLBCL, mais que leurs résultats cliniques en termes de SG et de survie sans progression (SSP) sont totalement opposés. Nous avons montré que ces variations sont principalement dues à des localisations tumorales plus agressives pour le CS3 et à un MET chaude pour le CS4. En ce qui concerne les cibles thérapeutiques potentielles, même si les deux groupes pourraient être potentiellement plus sensibles aux inhibiteurs de BTK (par exemple, l'ibrutinib), le groupe CS4 pourrait également bénéficier d'inhibiteurs de JAK1 et de points de contrôle immunitaire, soit parce qu'il présente une activité transcriptionnelle JAK-STAT élevée, soit parce qu'il présente une expression élevée du CMH-I en l'absence de délétions bialélique HLA. En outre, le groupe CS3 pourrait également bénéficier des ICI, mais seulement après exposition aux inhibiteurs d'EZH2, car il pourrait restaurer son expression manquante du CMH-I.

Il est intéressant de noter que les sous-types CS1 et CS2 du LPSNC étaient largement hyperméthylés par rapport aux autres, ce qui a été précédemment associé à un MET froid, comme observé au niveau transcriptionnel. Pour le CS1, la forte activité du complexe PRC2 et la prolifération (induite par des altérations génétiques impliquées dans le cycle cellulaire) ont été directement "vues" dans son phénotype hyperméthylateur. D'autre part, les programmes de différenciation des cellules B "perturbés" observés au niveau transcriptionnel chez les CS2 ont été corroborés au niveau épigénétique. Nous avons proposé que le groupe CS1 de froid immunitaire pourrait être sensible aux cyclines CDK4 et CDK6 ainsi qu'aux inhibiteurs de PI3K, tandis que le groupe CS2 pourrait être potentiellement sensible à l'inhibition des facteurs de transcription IRF4 (par exemple, lénalidomide), SPIB et MEIS1 (par exemple, MEISi-1), et/ou à l'inhibition de GAD67.

En ce qui concerne l'activité hors cible de c-AID, même si nous n'avons pas observé de différence dans le nombre global de mutations c-AID entre les sous-types

moléculaires, nous avons montré que globalement (en utilisant toute la cohorte de LPSNC) les mutations c-AID et non c-AID (signature cosmique SBS9) se produisent à des stades précoces de la tumorigenèse du LPSNC, ce qui reflète son importance dans la pathogenèse du LPSNC.

Puisque l'acquisition de tissus fraîchement congelés (FC) pour le LPSNC n'est pas effectuée de manière routinière dans les cliniques, nous avons validé nos résultats dans une cohorte de tissus fixés au formol et inclus en paraffine (FFPE en anglais) supplémentaire. De plus, nous avons développé RBraLymP (RNA-based Brain Lymphoma Profiler), qui utilise les données d'expression génique des tissus FFPE ou FC, pour identifier les sous-types moléculaires du LPSNC associés à des caractéristiques multi-omiques. Nous avons rendu le code accessible au public pour inciter les chercheurs du monde entier à orienter les efforts de nouvelles thérapies vers les patients atteints de LPSNC les plus appropriés.

Conclusion

La compréhension de l'hétérogénéité de la réponse moléculaire et clinique dans le LPSNC n'a pas été abordée de manière adéquate puisqu'elle a été construite sur la classification actuelle de LDGCB qui comprenait un faible nombre d'échantillons de LPSNC. Les résultats collectifs de ma thèse comblent cette lacune en reliant les caractéristiques multi-omiques intégrées dans chaque sous-type moléculaire de LPSNC à des cibles thérapeutiques potentielles. De plus, l'algorithme basé sur l'ARN, RBraLymP, peut faciliter les efforts futurs pour développer et évaluer des approches thérapeutiques ciblées pour ces malignités mal comprises et très meurtrières.

université PARIS-SACLAY | ÉCOLE DOCTORALE Cancérologie : biologie - médecine - santé (CBMS)

Titre : Analyse multi-omique des lymphomes primitifs du système nerveux central

Mots clés : Lymphomes primitifs du système nerveux central, Susceptibilité génétique, Cytidine désaminase induite par l'activation, Multi-omiques, Sous-types moléculaires

Résumé : Le lymphome primitif du système nerveux central (LPSNC) est un sous-type rare de lymphome diffus à grandes cellules B (LDGCB) situé dans le SNC avec un pronostic moins favorable. De plus, 60% des patients qui finiront par présenter une récidive après le traitement de référence (régime de méthotrexate à haute dose) ont montré des réponses hétérogènes dans divers essais cliniques avec différentes stratégies de traitement. En outre, la cytidine désaminase induite par l'activation (AID), qui déclenche physiologiquement l'hypermutation somatique et la recombinaison de changement de classe dans les cellules B du centre germinal, a une activité mutagène hors cible plus élevée dans le LPSNC par rapport aux autres LDGCB. Aujourd'hui, l'hétérogénéité du LPSNC. n'a pas été correctement abordée principalement en raison du manque d'intégration de données multi-omiques et du nombre limité de patients. Cette thèse se concentre sur la caractérisation et l'intégration du paysage multi-omique du LPSNC pour trouver des sous-groupes moléculaires du LPSNC ayant une pertinence biologique et clinique. Tout d'abord, nous avons examiné la structure HLA et la susceptibilité génétique dans le LPSNC et d'autres lymphomes à cellules B. Deuxièmement, nous avons développé et validé un code bioinformatique pour identifier les mutations AID et explorer leurs implications dans les lymphomes à cellules B et d'autres cancers. Enfin, nous avons intégré des données multi-omiques pour délimiter quatre classes moléculaires de LPSNC avec un impact pronostique remarquable et développé un algorithme qui utilise des données d'expression génique provenant de tissus fixés au formol et inclus en paraffine ou fraîchement congelés, pour identifier ces sous-types moléculaires de LPSNC. Collectivement, les résultats de cette thèse donnent des explications plausibles sur l'hétérogénéité de la réponse du LPSNC en trouvant un pont entre les différentes couches multi-omiques et les cibles thérapeutiques potentielles à travers des sous-types moléculaires du LPSNC.

Title: Multi-omics analysis of primary central nervous system lymphoma

Keywords: Primary central nervous system lymphoma, Genetic susceptibility, Activation-induced cytidine deaminase, Multi-omics, Molecular subtypes

Abstract: Primary central nervous system lymphoma (PCNSL) is a rare subtype of diffuse large B cell lymphoma (DLBCL) located in the CNS with a less favorable prognosis. Moreover, 60% of the patients that will eventually relapse from the gold standard treatment (high-dose methotrexate regimen), have shown heterogeneous responses in diverse clinical trials with different treatment strategies. Furthermore, the activation-induced cytidine deaminase (AID), which physiologically triggers somatic hypermutation and class-switch recombination in germinal-center B-cells, has a higher off-target mutagenic activity in PCNSL compared to other DLBCL. Today, PCNSL heterogeneity has not been properly addressed mainly due to the lack of multi-omic data integration and the limited number of patients. This thesis is focused on characterizing and integrating the multi-omic landscape of PCNSL to find molecular PCNSL subgroups with biological and clinical relevance. Firstly, we reviewed the HLA structure and the genetic susceptibility in PCNSL and other B-cell lymphomas. Secondly, we developed and validated a bioinformatic code to identify AID mutations and to explore their implications in B-cell lymphomas and other cancers. Finally, we integrated multi-omic data to delineate four molecular classes of PCNSL with a remarkable prognostic impact and developed an algorithm that uses gene expression data from either formalin-fixed, paraffin-embedded, or fresh-frozen tissue, to identify such PCNSL molecular subtypes. Collectively the findings in this thesis give plausible explanations on the PCNSL response heterogeneity based on finding a bridge between the different multi-omic layers and ultimately potential therapeutic targets across molecular PCNSL subtypes.