DISSERTATION

A novel methodology to identify DNA repair deficiencies by means of genomic scars generated in the course of antibody diversification

Eine neuartige Methode zur Identifizierung von DNA-Reparaturdefiziten anhand von genomischen Narben, die im Verlauf der Antikörperdiversifizierung auftreten

zur Erlangung des akademischen Grades

Doctor of Philosophy (Ph.D.)

vorgelegt der Medizinischen Fakultät

Charité – Universitätsmedizin Berlin

von

Clara Vázquez García

Erstbetreuer*in: Prof. Dr. Kathrin de la Rosa

Datum der Promotion: 30th June 2024

Table of Contents

1. ABBREVIATIONS

Abbreviation	Description
µL	Microliter
µM	micromolar
µm	Micrometer
2D-bps plot	Two-dimensional breakpoint plot
3´RR	3´regulatory region of IGH
53BP1	P53 Binding Protein 1
A	Adenine
aEJ	Alternative end-joining
AF	AlexaFluor
AID	Activation-induced cytidine deaminase
APE1	Apurinic-apyrimidinic endonuclease 1
ARPC1B	Actin Related Protein 2/3 Complex Subunit 1B
ATM	Ataxia-telangiectasia mutated kinase
ATR	Ataxia telangiectasia and Rad3 related
BACH2	Broad complex-tramtrack-bric a brac and Cap'n'collar homology 2
BAFF	B-cell activating factor
BARD1	BRCA1-associated RING domain protein 1
BCL-2	B cell lymphoma 2
BCR	B cell receptor
BER	Base excision repair
BIM	B cell lymphoma 2–interacting mediator
BLM	3′−5′ ATP-dependent RecQ DNA helicase
BM	Bone marrow
BRCA1	Breast cancer gene 1
BRCA2	Breast cancer gene 2
C	Cytosine
Cas9	CRISPR associated protein 9
CCL13	C-C chemokine ligand 13
CCL21	C-C chemokine ligand 21
CCR6	C-C chemokine receptor 6
CCR7	C-C chemokine receptor type 7
CD1b	Transmembrane glycoprotein from the CD1 family
CD21	Cluster of differentiation 21
CD23	Low affinity receptor for IgE
CD27	T-Cell Activation Antigen CD27
CD40	Cluster of differentiation 40
CD40L	CD40 ligand
CD83	Cluster of Differentiation 83
CD86	Cluster of Differentiation 86
CDR1	Complementarity Determining Region 1

CDR2	Complementarity Determining Region 2
CDR3	Complementarity Determining Region 3
CH	Constant heavy chain gene
CH12	Mouse B cell lymphoma cells
chr	Chromosome
CK2	Casein kinase 2
C_M	Constant μ
cNHEJ	Classical non-homologous end-joining
CRISPR	Clustered regularly interspaced short palindromic repeats
CSR	Class switch recombination
CST	CTC1-STN1-TEN1
CTC1	CTS Telomere Maintenance Complex Component 1
CTCF	CCCTC-binding factor
CtIP	CtBP-interacting protein
CVID	Common variable immunodeficiency
CXCL10	C-X-C motif chemokine ligand 10
CXCL11	C-X-C motif chemokine ligand 11
CXCL12	C-X-C motif chemokine ligand 12
CXCL13	C-X-C motif chemokine ligand 13
CXCL9	C-X-C motif chemokine ligand 9
CXCR3	C-X-C Motif Chemokine Receptor 3
CXCR4	C-X-C Motif Chemokine Receptor 4
CXCR5	C-X-C Motif Chemokine Receptor 5
D	Diversity
DC	Dendritic cells
DNA	Deoxyribonucleic acid
DNA2	DNA Replication Helicase/Nuclease 2
DNA-PKcs	DNA-dependent protein kinase, catalytic subunit
DRB1	RNA debranching enzyme 1
DRdef	DNA repair deficient
dsDNA	double-stranded DNA
DZ	Dark zone of the germinal center
EBI2	Epstein-Barr virus-induced G-protein coupled receptor 2
EDTA	Ethylenediaminetetraacetic acid
eff_nclusters	The smallest number of nclusters containing 95% of the reads
ELF3	E74 Like ETS Transcription Factor 3
EM	Enhancer μ
ERCC1	Excision repair cross complementation group 1
ERFS	Early replication fragile sited
EXO1	Exonuclease 1
FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
FDCs	Follicular dendritic cells

FO	Follicular
FISH	Fluorescence *in situ* hybridization
FR1	Framework area 1
FR2	Framework area 2
FR3	Framework area 3
FR4	Framework area 4
FW	Forward
G	Guanine
g	Grams
GC	Germinal center
gDNA	Genomic DNA
GLTs	Germline transcripts
GRCh37	Genome Reference Consortium Human Build 37
GRCh38	Genome Reference Consortium Human Build 38
GRCm38	Genome Reference Consortium Mouse Build 38
GRCm39	Genome Reference Consortium Mouse Build 39
GWAS	Genome-wide association studies
Gy	Grades
H2AX	Histone family member X
H3	Histone 3
H4	Histone 4
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HD	Healthy donor
HeLa	Henrietta Lacks
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HIV	Human immunodeficiency virus
HMGB1	High mobility group B 1
HMGB2	High mobility group B 2
HR	Homologous recombination
HSPCs	Hematopoietic stem and progenitor cells
hTR	Human telomerase RNA
HU	Hydroxyurea
ICAM-1	Intercellular Adhesion Molecule 1
Ig	Immunoglobulin
IG1	Promoter of IGHG1
IgA	Immunoglobulin α
IgD	Immunoglobulin δ
IgE	Immunoglobulin ε
IgG	Immunoglobulin γ
IGH	Immunoglobulin heavy chain locus
Igh	Immunoglobulin heavy chain
IGHA1	Immunoglobulin heavy chain α1 locus

IGHA2	Immunoglobulin heavy chain α2 locus
IGHD	Immunoglobulin heavy chain δ locus
IGHE	Immunoglobulin heavy chain ε locus
IGHG1	Immunoglobulin heavy chain γ1 locus
IGHG2	Immunoglobulin heavy chain γ2 locus
IGHG3	Immunoglobulin heavy chain γ3 locus
IGHG4	Immunoglobulin heavy chain γ4 locus
IGHM	Immunoglobulin heavy chain μ locus
IGL	Immunoglobulin light chain locus
Igl	Immunoglobulin light chain
IGLK	Immunoglobulin light chain κ locus
IGLL	Immunoglobulin light chain λ locus
IgM	Immunoglobulin μ
IL2	Interleukin 2
IL4	Interleukin 4
IM	Promoter of IGHM
indel	Insertions and deletions
IRF4	Interferon regulatory factor 4
ITS	Interstitial telomeric sequence
J	Joining
JARID2	Jumonji And AT-Rich Interaction Domain Containing 2
J-CH1 inserts	Insertions found between the J segment and the CH1 exon
K	1 000
Kb	Kilobase (1 000 bases)
LAIR1	Leucocyte-associated immunoglobulin-like receptor 1
LAM-HTGTS	Linear amplification high-throughput genome-wide translocation sequencing
LFA-1	Lymphocyte function-associated antigen 1
LILRB1	Leucocyte immunoglobulin-like receptor B1
L-NGS	Long-read sequencing
LZ	Light zone of the germinal center
med.dist	Median distance
MH	Microhomology
MHC	Major histocompatibility complex
min	Minutes
ml	Milliliter
MMEJ	Microhomology-mediated end joining
Mre11	Meiotic Recombination 11
MRN	Mre11-RAD50-NBN
mRNA	messenger RNA
SBS9	Single base substitution signature 9
mSA	Mouse switch α region
mSG2bc	Mouse switch γ2bc region
mSG3	Mouse switch γ3 region

mSM	Mouse switch µ region
NMD	Non-sense-mediated decay
mtDNA	Mitochondrial DNA
mM	Milimolar
mTGF-b1	Mouse Transforming growth factor beta 1
MutSa	mismatch DNA repair protein type A
MYC	MYC Proto-Oncogene, BHLH Transcription Factor
MZ	Marginal zone
nc-BER	non-canonical base excision repair
nc-MMR	non-canonical mismatch repair
NFKB1	Nuclear Factor Kappa B Subunit 1
ng	nanograms
NIPBL	Nipped-B-like protein
NK	Natural killer
NMD	Non-sense mediated decay
nt	Nucleotide
NTD	N-terminal domain
PALB2	Partner and localizer of BRCA2
PARP1	Poly [ADP-ribose] polymerase 1
PBMCs	Human peripheral blood mononuclear cells
PBS-T	Phosphate-buffered saline solution
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PLK1	Polo-like kinase 1
Pol	Polymerase
POT1	Protection of telomeres protein 1
PP	Peyer´s patches
PRDM1	PR/SET Domain 1
R	Adenine or Guanine
RAD51	DNA repair protein RAD51 homolog 1
RAD52	DNA repair protein RAD52 homolog 2
RAG1	Recombination activating gene 1
RAG2	Recombination activating gene 2
Rif1	Replication timing regulatory factor 1
RIFIN	Repetitive interspersed families of polypeptides
RNA	Ribonucleic acid
RNR	Ribonucleotide reductase
ROS	Reactive oxygen species
RPA	Replication protein A
RPMI	Roswell Park Memorial Institute Medium
RSS	Recombination signal sequences
RV	Reverse

s	Second
SA	Switch α region
SA1	Switch α1 region
SA2	Switch α2 region
SE	Switch ε region
SG1	Switch γ1 region
SG2	Switch γ2 region
SG2b	Switch γ2b region
SG2c	Switch γ2c region
SG3	Switch γ3 region
SG4	Switch γ4 region
SHLD1	Shieldin complex subunit 1
SHM	Somatic hypermutation
SM	Switch μ region
SNP	Single nucleotide polymorphism
SSA	Single-strand annealing
ssDNA	single-stranded DNA
STN1	Oligosaccharide Binding Fold Containing 1
SWIBRID	SwItch region breakpoint repertoire identification
Switch-joint	Joint generated by class-switch recombination between two different switch regions
SX	Any switch region that is not SM
T	Thymine
T1	Transitional 1
T2	Transitional 2
TCR	T cell repertoire
TdT	Terminal deoxynucleotidyl transferase
TEN1	Telomere Length Regulation Protein 1
TERT	Telomerase Reverse Transcriptase
Tfh	T follicular helper
TPP1	Tripeptidyl-peptidase 1
TRF2	Telomeric repeat factor 2
U	Uracil
UNG	Uracil-DNA glycosylase
UV	Ultraviolet
V	Variable
VCAM-1	Vascular cell adhesion protein 1
VLA-4	very late antigen-4
W	Adenine or Thymine
WT	Wild-type
Xlf1	XRCC4-like factor 1
XPF	Xeroderma Pigmentosum Group F-Complementing Protein
XRCC1	X-ray repair cross-complementing protein 1
XRCC4	X-ray repair cross-complementing protein 4

Y	Cytosine or Thymine
ZFP318	Zinc-finger protein 318

2. LIST OF TABLES AND FIGURES

List of Tables:

List of Figures:

3. ABSTRACT

3.1. Abstract

I developed a methodology to study class-switch recombination (CSR) genomic scars, also called switch-joints, in B cell DNA. I applied this methodology to study a recently discovered antibody diversification process and the DNA repair malfunction in humans and mice.

CSR is the diversification process by which the antibody changes the isotype. It is a deletional recombination occurring in the antibody gene between long repetitive intronic sequences called switch regions. During CSR, activation-induced cytidine deaminase (AID) promotes random breaks in two switch regions, later joined by the main DNA repair pathways. Thus, the study of switch-joints in polyclonal B cells allows the analysis of diverse genomic scars. B cells can also diversify their antibody by a recently discovered process integrating foreign DNA fragments, called inserts, in the switch regions. Inserts have been analyzed in antibody transcripts, but the mechanism of their generation remains unknown.

This methodology, called SWIBRID, involves the amplification, long-read sequencing, and computational analysis of switch-joints. Here, I designed a PCR to amplify human and mouse switch-joints using as few as three primers. In addition, published switch-joint analyses use read count as a metric for clonal diversity. We demonstrated that the read count does not reflect the unique clone composition and used this insight to identify unique B cell clones by SWIBRID and confirmed its proper performance using an *in silico* dataset.

Firstly, SWIBRID was used to analyze the incorporation of inserts in switch-joints. Results of *in vitro* and *in vivo* modulation of classical and alternative end-joining DNA repair pathways in human samples suggested that alternative end joining is involved in insert acquisition. Also, AID off-targets telomeres and a similar case may occur with telomerase in the switch regions. Thus, I inhibited telomerase in *in vitro* activated human B cells and found its potential involvement in incorporating a specific insert type.

Secondly, we theorized that early detection of diseases like cancer might be aided by DNA repair malfunction analysis via SWIBRID. I studied DNA repair by analyzing genomic scars. While challenging in somatic cells due to the random occurrence of DNA breaks in the genome, B cells serve as a subject for this analysis since they acquire breaks during CSR. Four criteria were identified to describe the switch-joint genomic scars, e.g., homology used for repair. BRCA1, Ligase 4, NIPBL, ATM, and AID deficient human patients and BRCA1, 53BP1, Rif1 and Ligase 4 mouse CH12 knockouts were analyzed using a principal component analysis (PCA). Using only those features, the PCA separated the DNA repair deficient from the healthy samples. Results show that healthy- and disease-specific breakpoint profiles may serve as a DNA repair biomarker. In the future, this pipeline could be used to predict DNA repair deficiencies linked to cancer development or immunodeficiencies.

3.2. Zusammenfassung

In dieser Arbeit habe ich eine Methode entwickelt, um durch Class-Switch-Rekombination (CSR) erzeugte DNA Narben (Switch-Joints), in B-Zellen zu analysieren, um die Antikörper-Diversifizierung sowie DNA-Reparaturfehlfunktionen zu erforschen.

CSR beschreibt den Prozess, durch den ein Antikörper sein Isotyp ändert. Es handelt sich dabei um eine deletionale Rekombination zwischen langen, repetitiven intronischen Sequenzen (Switch-Regionen) im Antikörper-Gen. Während der CSR fördert die activation-induced cytidine deaminase (AID) zufällige Brüche in zwei Switch-Regionen, die später durch DNA-Reparatur verbunden werden. Switch-Joints in polyklonalen B-Zellen eignen sich daher sehr für die Analyse von DNA-Narben. B-Zellen können ihre Antikörper auch durch einen anderen Prozess diversifizieren, bei dem fremde DNA-Fragmente (Inserts) in die Switch-Regionen integriert werden. Der Mechanismus ihrer Entstehung ist aber noch unbekannt.

Die von mir entwickelte Methode heißt SWIBRID und umfasst die Amplifikation mit drei Primern, Long-Read-Sequenzierung und computergestützte Analyse von humanen sowie murinen Switch-Joints. Entgegen bislang publizierter Studien konnten wir zeigen, dass nicht die Anzahl der Reads der wahren Zusammensetzung der Klone entspricht und mit Hilfe von SWIBRID eindeutige B-Zell-Klone identifizieren.

Zuerst wurde SWIBRID verwendet, um die Integration von Inserts in Switch-Joints zu analysieren. Die Ergebnisse der *in vitro*- und *in vivo*-Modulation der klassischen und alternativen End Joining-DNA-Reparatur in humanen Proben legten nahe, dass Letztere an der Integration beteiligt ist. Außerdem wurden Telomere als eines der Off-Targets von AID beschrieben, und ein ähnliches Phänomen könnte auch bei der Telomerase in den Switch-Regionen auftreten. *In vitro* Inhibition der Telomerase in aktivierten humanen B-Zellen zeigte, dass sie für den Einbau einer bestimmten Art von Insert verantwortlich sein könnte.

Weiterhin stellte ich die Hypothese auf, dass SWIBRID für die Früherkennung von Krebs durch die Analyse von Fehlfunktionen der DNA-Reparatur genutzt werden könnte. B-Zellen eignen sich im Vergleich zu somatischen Zellen besonders gut für diese Analyse, da sie während der CSR, gerichtete DNA-Brüche erleiden. Basierend auf vier ausgewählten Kriterien wurden DNA-Narben in BRCA1, Ligase 4, NIPBL, ATM und AID defizienten humanen Proben sowie in BRCA1, 53BP1, Rif1 und Ligase 4 murinen CH12 Knock-out Zellen mit einer Hauptkomponentenanalyse (PCA) analysiert. Die PCA zeigte, dass die vier Merkmale ausreichten, um die Proben mit DNA-Reparatur Defiziten von den gesunden Proben zu unterscheiden. Die Ergebnisse deuten darauf hin, dass es gesunde und krankheitsspezifische Bruchpunktprofile gibt, die als DNA-Reparatur-Biomarker dienen könnten. In Zukunft könnte diese Pipeline zur Vorhersage von Fehlern in der DNA-Reparatur im Zusammenhang mit Krebs oder Immundefekten genutzt werden.

4. INTRODUCTION

4.1. B cell biology

The adaptive immune system, comprised of T and B cells, can efficiently eliminate threats by specifically recognizing pathogens and building memory. T cells are responsible for cell-mediated immunity, while B cells are responsible for antibody-mediated immunity. B cells produce membrane-bound and secreted antibodies. When the antibody is expressed in the membrane, it is known as the B cell receptor (BCR). In B cells, the cell maturation and the antibody diversification processes go hand-in-hand since B cells are classified into distinct populations defined by the diversification stage of the BCR. In brief, B cells are developed from hematopoietic precursors cells in the bone marrow (BM) until they egress as immature B cells when their BCR is ready to bind foreign antigens. Outside the BM and upon antigen encounter, they continue diversifying in lymph nodes into other B cell subpopulations, such as plasmablast or memory B cells (*1*).

The antibody, which is part of the immunoglobulin (Ig) superfamily, is a protein that targets threatening epitopes, a fragment of an antigen, and promotes their elimination by distinct pathways. The antibody comprises 2 immunoglobulin heavy chains (Igh) and 2 Ig light chains (Igl) (Figure *1*A). The Ighs are bound to each other by disulfide bonds, as well as each Igh to an Igl. Both chains contain a variable and a constant region, the former determines specificity, while the latter determines the isotype of the Ig. In humans, the variable region in the immunoglobulin heavy chain locus (IGH) contains 159 variable (V) segments, 27 diversity (D) segments and 9 junction (J) segments, and the constant region contains exon clusters organized in the order of IGH μ (IGHM), IGH δ (IGHD), IGH γ3 (IGHG3), IGH γ1 (IGHG1), IGH α1 (IGHA1), IGH γ2 (IGHG2), IGH γ4 (IGHG4), IGH ε (IGHE) and IGH α2 (IGHA2) and are encoded in chromosome 14. Each Ig exon cluster contains different numbers of exons, called heavy constant genes (CH). The variable and CH1 are separated from the CH2 in the protein by a flexible hinge region. Compared to the Igh, the two distinct Igl are encoded on two chromosomes, and its variable region does not contain D domains. In chromosome 2 is immunoglobulin light chain locus (IGL) κ (IGLK), which variable region contains 77 V segments and 5 J segments and a constant region carrying IGLK. In chromosome 22 is IGL λ (IGLL), which variable region contains 80 V segments and 9 J segments and a constant region carrying IGLL (*2*) (Figure *1*B).

IGH and IGL genes undergo diversification processes that allow the generation of up to 10^{15} combinations of antibodies, as defined by the recombined V(D)J recombination (*3*). The amount of diverse antibodies in an individual determines the B cell repertoire (*4*, *5*). The diversification processes are called i) V(D)J recombination, which occurs in the variable region of IGH and IGL (*6*); ii) class-switch recombination (CSR), which occurs in the constant region of the IGH and iii) somatic hypermutation (SHM) which occurs in the variable and constant regions of the IGH and IGL genes (*7*). V(D)J recombination is carried out in the

BM by recombination activating gene 1 (RAG1) and 2 (RAG2) proteins, allowing the BCR to be expressed carrying a V(D)J recombined domain for the B cells to egress the BM towards the secondary lymphoid organs. CSR and SHM are promoted by activation-induced cytidine deaminase (AID) and are active during the development of a specific immune response in the germinal center (GC) (section 4.3.4) (*8*).

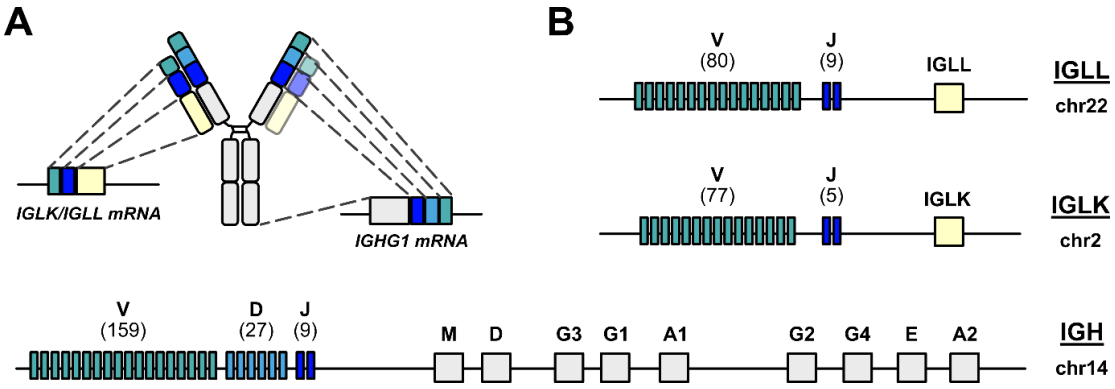


Figure 1. Human IGH, IGLL and IGLK genes and their contribution to the antibody.

A. Representation of the antibody. Location of the IGH and IGL genes in the antibody. **B.** Human IGH, IGLL and IGLK genes are divided into variable (V, D, J) and constant regions (in yellow and grey). mRNA=messenger RNA, IGH=Immunoglobulin heavy chain, IGL=Immunoglobulin light chain, chr=chromosome, IGLL=IGL λ, IGLK=IGL κ, M=IGHM, D(grey)=IGHD, G3=IGHG3, G1=IGHG1, A1=IGHA1, G2=IGHG2, G4=IGHG4, E=IGHE, A2=IGHA2, V=Variable, D(blue)=Diversity, J=Joining. In brackets, the number of different V, D and J domains in the human variable region of the IGH and IGLs.

4.2. B cell development

B cell diversification starts in the BM, where B cells maturate until they egress to the periphery. In the following explanation, I will introduce the process of B cell maturation in humans. Firstly, the hematopoietic precursor cells rearrange the D-J segments and start expressing a pro-BCR, becoming pro-B cells (*9*). Subsequentially, they rearrange the V to the DJ segments in the IGH and become pre-B cells expressing IgM carrying a surrogate light chain (*10*). Pre-B cells rearrange V and J segments of the IGL and express Ig IGHM (IgM), becoming immature B cells. Immature B cells can leave the BM and enter the blood circulation to encounter an antigen triggering further B cell maturation processes in the GC to become plasmablast, plasmacells or memory B cells (*11*).

Immature B cells egressing the BM migrate to lymphoid organs such as the spleen. During the transition from the BM to the lymphoid organ, B cells are called transitional 1 (T1) B cells and are recognized by the markers IgD^-

$IgM^{+}CD21^{LOW}CD23^{-}$. The cluster of differentiation 21 (CD21) and the low receptor for IgE (CD23) are markers that provide survival signals to B cells not yet positively selected (*12*, *13*). T1 B cells maturate to transitional 2 (T2) B cells upon B cell activation factor (BAFF) signaling, promoted by myeloid cells in the lymphoid organ (*14*). T2 B cells start expressing Ig IGHD (IgD) in their membrane. Co-expression of IgM and IgD isotypes is the result of BCR splice variants carrying the same rearranged V(D)J but different constant Igh. IgD expression is aided by the zinc-finger protein 318 (ZFP318), expressed in T2 B cells, that allows the ribonucleic acid (RNA) polymerase II read-through of the stop codon in IGHM, transcribing the IGHD exon cluster. Generally, IGHM transcription has been observed to be 3-fold higher than IGHD in T2 B cells (*15*, *16*). T2 B cells are recognized by the markers $IgD^{HI}IgM^{HI}CD21^{MID}CD23^{+}$ and can also be called naïve B cells. T2 B cells are a transitional stage of B cells whose fate depends on the BCR binding strength to the antigen. If the strength is high, they will become follicular (FO) B cells; if low, they will become marginal zone (MZ) B cells. FO B cells migrate to the T cell zone of the lymph node and are the ones that likely enter the GC for further maturation. FO B cells can be recognized by the markers $IgD^{HI}IgM^{+}CD21^{MID}CD23^{+}$. On the other hand, MZ B cells are located in the outer white pulp of the spleen. MZ B cells are the first encounters of immunocomplexes, and their high expression of CD21 has been linked to the transport of antigens to the GC. MZ B cells can be recognized by the markers $IgD^{LOW}IgM^{HI}CD21^{HI}CD23^{-}CD1^{HI}$ (*17*).

FO B cells would encounter an antigen and start expressing Epstein-Barr virus-induced G-protein coupled receptor 2 (EBI2) and C-C chemokine receptor type 7 (CCR7). EBI2 is vital to keep FO B cells in the outer part of the follicle where the GC is carried out (*18*). On the other hand, CCR7 leads the FO B cells to migrate towards the T:B border driven by residing stromal cells that highly express C-C chemokine ligand 21 (CCL21), the CCR7 ligand. In the T:B cell border, there are T follicular helper (Tfh) cells that have been previously activated by dendritic cells (DC) (*1*, *19*). Tfh cells test if the FO B cells arriving at the border carry BCRs binding to the antigens presented in the GC. In order to do so, FO B cells internalize the antigen encountered, process it and present fragments of it on the major histocompatibility complex (MHC) class II to the Tfh cells (*20*). The antigen fragment is presented to the T cell receptor (TCR) via the MHC class II, facilitated by a synapsis formed by the CD40 ligand (CD40L) and the cluster of differentiation 40 (CD40) between Tfh cells and FO B cells, respectively (Figure *2*). When Tfh cells recognize the fragment via their TCR, they secrete cytokines, e.g., interleukin 4 (IL4), which serve the FO B cells as pro-survival signals. These signals push the downregulation of the BCR and the upregulation of AID and the chemokine receptor 4 (CXCR4) and 5 (CXCR5). These changes in expression transform the FO B cell into a GC B cell. Nonetheless, a subset of activated FO B cells that intrinsically has a higher affinity towards the antigen would differentiate into short-lived plasmablasts. The short-lived plasmablasts are part of a short-timed adaptive response and secrete large amounts of antibodies to combat ongoing threats (*21*, *22*).

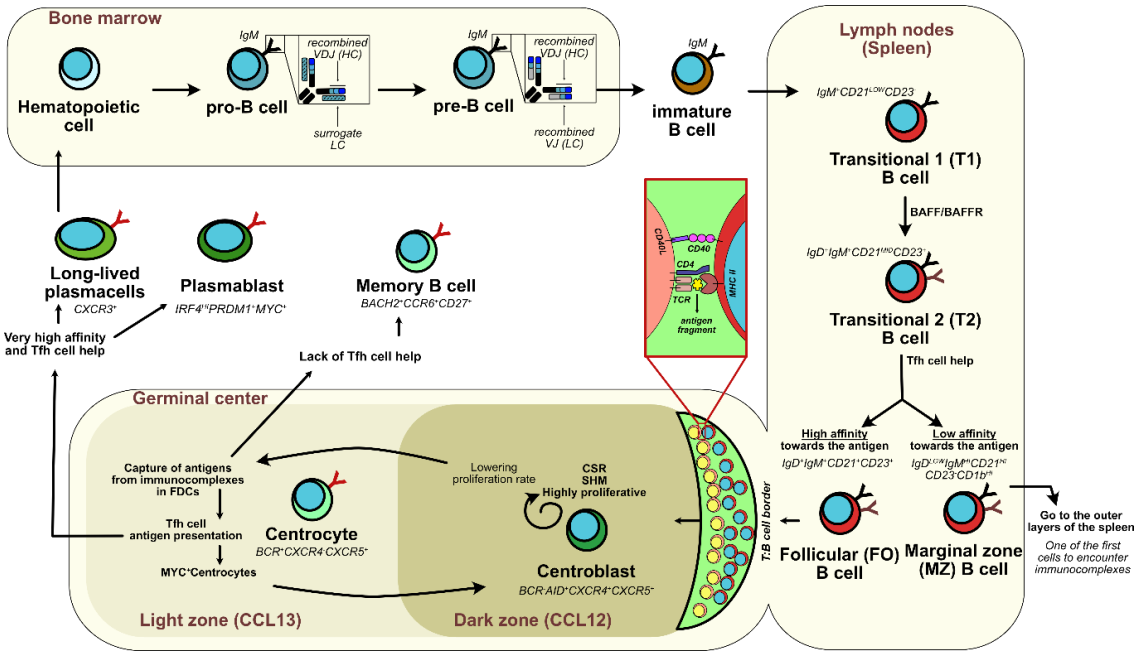


Figure 2. B cell development.

Scheme representing the most relevant B cell populations during their development. Boxes represent the different anatomical compartments where the development occurs: bone marrow and lymph nodes. The germinal center is formed inside the lymph nodes. VDJ=recombined V, D and J domains of the immunoglobulin heavy chain, HC=heavy chain, LC=light chain, VJ=recombined V and J domains of the immunoglobulin light chain, IgM=Immunoglobulin μ, CD21=cluster of differentiation 21, CD23=low-affinity receptor for IgE, BAFF=B cell activating factor, BAFFR=BAFF-receptor, IgD=Immunoglobulin δ, CD1b=transmembrane glycoprotein from the CD1 family, BCR=B cell receptor, CSR=Class switch recombination, SHM=somatic hypermutation, CXCR4=C-X-C motif chemokine receptor 4, CXCR5=C-X-C motif chemokine receptor 5, CCL12=C-C chemokine ligand 12, CCL13=C-C chemokine ligand 13, AID=Activation-induced cytidine deaminase, FDCs=follicular dendritic cells, CXCR3=C-X-C motif chemokine receptor 3, IRF4=interferon regulatory factor 4, PRDM1=PR/SET Domain 1, MYC=MYC Proto-Oncogene, BACH2=broad complex-tramtrack-bric a brac and Cap'n'collar homology 2, CD27=T-Cell Activation Antigen CD27, CCR6=C-C Motif Chemokine Receptor 6, TCR=T cell receptor, MHC II=Major histocompatibility complex class two, CD4=CD4 mo, CD40=Cluster of differentiation 40 , CD40L=CD40 ligand.

The GC is a microanatomical organization in which the maturation of the B cells is carried out through the orchestration of GC B cells, Tfh cells and follicular dendritic cells (FDCs). The GC can be located in secondary lymphoid organs and is divided into the light zone (LZ) and the dark zone (DZ). GC B cells in the DZ, called centroblasts, are recognized by $CXCR5^{+}CXCR4^{HI}CD83^{LOW}CD86^{LOW}$ and in the LZ, called centrocytes, are recognized by $CXCR5^{+}CXCR4^{LOW}CD83^{HI}CD86^{HI}$. In contrast, Tfh cells in the GC express CXCR5 in the LZ and CXCR4 in the DZ (*23*). The cluster of differentiation 83 (CD83) and 86 (CD86) expression has been linked to the expression of MHC

class II (*24*), which agrees with their expression during LZ to present their antigen fragments to the Tfh cells. The LZ is home to stromal cells expressing C-X-C motif chemokine ligand 13 (CXCL13), the ligand of CXCR5, while the DZ is home to stromal cells expressing the C-X-C motif chemokine ligand 12 (CXCL12), the ligand of CXCR4. As previously described, when FO B cells become GC B cells, they downregulate the BCR and up-regulate AID, CXCR4 and CXCR5. Thus, freshly activated GC B cells migrate to the DZ and undergo CSR and SHM, promoted by the high AID expression and acquire mutations in their BCRs to improve the affinity towards the antigen requiring extensive proliferation and to class-switch the Ig isotype to acquire different effector functions (*25*). However, mutations are also detrimental, in which case there are mechanisms to control the consequences explained in section 4.3.4. The centroblasts that slow down the proliferation also stop expressing CXCR4, forcing the centroblast to migrate to the LZ and become centrocytes (Figure *2*). FDCs in the LZ present the native antigen that drives B cell maturation. They attract $CXCR5^{+}$B cells with CXCL13 membrane receptors. B cells attach to the FDCs via the lymphocyte function-associated antigen 1 (LFA-1) to the intercellular adhesion molecule 1 (ICAM-1) and the very late antigen-4 (VLA4) to the vascular cell adhesion protein 1 (VCAM-1) interaction. This interaction is necessary for the centrocytes to access the native antigens presented by the FDC and to receive pro-survival signals through BAFF recognition (*26–28*). Successful antigen binding is followed by antigen internalization and fragment presentation in the MHC class II. Again, B cells pursue pro-survival signals through interaction with Tfh cells. If the centrocyte receives pro-survival signals from the Tfh cells, the MYC proto-oncogene (MYC) gene would be expressed, which marks the centrocytes that have received positive signals to continue cycling through the GC reaction (*29*).

B cells cycle between the LZ and DZ until plasmablast, plasmacells and memory B cells eventually egress the GC to the periphery. Centrocytes with a very high affinity towards an antigen internalize more antigens and can present more fragments in the MHC class II to the Tfh cells. Tfh cells create a more prolonged contact with them due to the higher MHC class II presentation and secrete interleukin 2 (IL2) and IL4 (*30*). Even though not all these B cells become plasmablasts, this has been described as the path for differentiation into plasmablast. A small set of these plasmablasts express chemokine receptor 3 (CXCR3) and become long-lived plasmacells, migrating to the BM where the CXCR3 ligands, chemokine ligand 9 (CXCL9), 10 (CXCL10) and 11 (CXCL11), are expressed (*27*, *31*). On the other hand, memory B cells arise from centrocytes that were not positively selected by Tfh cells, thus, from low-affinity B cells. This assumption was made since memory B cell precursors express broad complex-tramtrack-bric a brac and Cap'n'collar homology 2 (BACH2), a gene downregulated upon Tfh cell help. Memory B cell precursors pause their cell cycle and start expressing CCR6, BACH2 and CD27, a TNF family receptor. They migrate from the LZ to an intermediate GC stage before migrating to the periphery (*27*, *31–34*) (Figure *2*).

The homeostasis in the GC is guided by B cell apoptosis. In fact, half of the GC B cells die every six hours (*35*). In the LZ, GC B cells die when not receiving any

Tfh cell help, thus not promoting MYC expression and entering apoptosis due to Jumonji And AT-Rich Interaction Domain Containing 2 (JARID2) accumulation in the nucleus that impedes the progression of the cell to the S phase (*35*, *36*). In the DZ, GC B cells die after the diversification processes when their BCR is damaged by mutations that hinder their expression. This event has been observed in around 60% of the centroblasts (*35*).

B cell populations in the blood represent the frequency of B cell populations in the lymphoid organs, except for GC B cells. The primary group of cells in the blood comprises erythrocytes and peripheral blood mononuclear cells (PBMCs). B cells ($CD19^+$) represent around 8% of PBMCs in the blood. Specific markers were used to estimate the populations of B cells present in healthy blood, i) 20-42% memory B cell ($CD27^+$), ii) 3-7% plasmablasts ($IgD^-CD27^{HI}CD38^{MID}CD138^-$), iii) 0.3% plasmacells ($IgD^-CD27^{HI}CD38^{MID}CD138^+$) and iv) 49-52% naïve B cells ($CD27^-IgM^+IgD^+$) (*37*).

4.3. B cell diversification – VDJ, SHM and CSR

As previously described, antibody diversity starts in the BM when random and imprecise genetic recombination of V, D, and J domains creates a vast number of antibodies with unique binding sites. Following the V(D)J recombination, mature B cells leave the BM and get activated upon antigen encounter in the periphery, starting the process of CSR and SHM. Both processes further diversify the antibody repertoire: CSR by introducing large genomic deletions that induce switching from IgM to other immunoglobulin classes, namely IgD, IgG, IgA, or IgE, and SHM by introducing mutations in the recombined V(D)J domains to improve antigen affinity (*38*, *39*).

4.3.1. VDJ recombination

The V(D)J recombination is a process that irreversibly recombines V, D and J domains in the variable regions of IGH and IGL genes to generate the antibody binding domain (*40*). It is known as V(D)J recombination because, in the IGH, there are V, D and J domains, while in the IGL, there are only V and J domains. This process starts in early pro-B cells and finalizes in pre-B cells. This process occurs in the BM and is carried out most importantly by RAG1 and RAG2 because RAG2 is necessary for RAG1 to function, which is to nick the V, D, and J segments. A protein complex comprising the RAG proteins recognizes conserved deoxyribonucleic acid (DNA) motifs adjacent to every individual V, D and J segment called recombination signal sequences (RSS). The RSS consists of a conserved heptamer (5´-CACAGTG-3') spaced by 12 or 23 nt to a conserved nonamer (5´-ACAAAAACC-3'). RAG recombination only occurs between two protein complexes in which one carries an RSS spaced by 12 nt, and another carries an RSS spaced by 23 nt. This is known as the 12/23 rule. The 12/23 rule assures that the V segment cannot be recombined to a J segment (23 nt spacers) without a D segment (12 nt spacers) in the IGH (*41*). The V(D)J recombination in the IGH occurs firstly between the D-to-J segments in early pro-B cells, which

usually happens in both alleles, and then, late pro-B cells undergo V-to-DJ recombination to give rise to pre-B cells (*1*, *40*).

Firstly, RAG1 and RAG2 must be available for V(D)J recombination. RAG1 is expressed throughout the entire cell cycle and aggregates in the nucleoli by its N-terminal domain (NTD). RAG1 aggregation is part of the posttranslational regulation of the V(D)J recombination, which impedes RAG1 from functioning during the whole cell cycle (*42*). On the other hand, RAG2 expression is tightly regulated by the cell cycle proteins. RAG2 is degraded constantly except during the G1 phase, exactly when V(D)J recombination occurs. The recombination occurs only during G1 because the classical non-homologous end-joining (cNHEJ) is the major DNA repair process during this cell phase. cNHEJ is an error-prone DNA repair pathway crucial to generate diversification during the V(D)J recombination (*43*). Thus, the function of RAG2 is to disaggregate RAG1 in the nucleoli and start with the V(D)J recombination by forming a protein complex (*44*).

Secondly, the protein complex must be formed, recognize the RSS, and nick the V, D and J domains for further recombination. The protein complex is formed by RAG1, RAG2, High mobility group B 1 (HMGB1) and 2 (HMGB2). The last two proteins promote DNA bending for enhanced synapsis and cleavage. Two protein complexes in parallel scan for RSS in the sequences. When the RAG1 in two independent protein complexes finds the RSS, a paired complex will be formed by joining the two protein complexes. The pair complex only happens following the 12/23 rule; in other words, between a protein complex that recognizes a nonamer of an RSS with a 12 nt spacer and the other recognizes a nonamer with a 23 nt spacer (*6*, *45*). The pair complex is necessary to bring the two segments to be recombined together (D-J, V-J or V-DJ). Then, RAG1 nicks the DNA between the RSS and the coding sequences, forming a hairpin on the coding side and a blunt end on the RSS side.

Thirdly, the hairpin needs to be resolved, and the two double-strand (ds) DNA breaks need to be ligated to each other. At this point, cNHEJ proteins are crucial for joining the domains. The heterodimer Ku70/Ku80 binds the DNA ends to protect them from further degradation. DNA-dependent protein kinase, catalytic subunit (DNA-PKcs), couple with Ku70/Ku80 in the DNA ends. DNA-PKcs and Artemis form a complex, and DNA-PKcs phosphorylates Artemis to acquire endonucleolytic 5' and 3' activity. The endonucleolytic activity allows Artemis to open the hairpins formed after the RAG1 nick. Artemis cuts the hairpin with high variability and forms ends with 2-8 nt of single-stranded (ss) DNA extensions for further processing (*46*, *47*) (Figure *3*A). When opening a hairpin, the DNA acquires a palindromic sequence. The open coding ends are potential substrates for terminal deoxynucleotidyl transferase (TdT). TdT is a polymerase that extends the open ends with non-templated, or random, nucleotides (*48*, *49*). After adding non-templated nucleotides, the cNHEJ joins the two ends using blunt end ligation or less than 2 nt homology (section 4.5.1).

The V(D)J recombination gives rise to the antibody binding domain comprised of three hypervariable regions: complementarity-determining regions 1 (CDR1), 2

(CDR2) and 3 (CDR3), and four framework areas: FR1, FR2, FR3 and FR4. The CDRs form the antibody binding domain, which conformation is maintained by the FDRs. CDR1 and CDR2 are coded by the V domain, while the recombination of V, D and J in the IGH and V and J in the IGL codes CDR3. It is assumed that CDR3 is the most diverse CDR of all three since it arises from the recombination of three segments. Further diversification of the VDJ domain is provided by SHM (section 4.3.3) (*50*–*53*).

4.3.2. Class switch recombination

CSR is the process by which B cells exchange their Ig class, providing antibodies with different effector functions (Table *1*). CSR occurs in the constant part of the IGH. It is an irreversible join between the introns preceding the exons in the IGH, also referred to as deletional recombination. The CSR is carried out in the so-called switch regions, characterized by long, highly repetitive intronic GC-rich sequences. In brief, CSR starts with double-stranded DNA (dsDNA) breaks in two switch regions promoted by AID, the first being the switch μ region (SM) and the second any other switch region downstream. These breaks come together to be repaired primarily by cNHEJ. The most critical mechanistic players in CSR are germline transcripts (GLTs), AID and loop extrusion.

Every Ig isotype gene comprises a cytokine-inducible promoter, an intervening exon, a switch region and a constant gene comprised of a cluster of exons. The Ig transcription and subsequent splicing generate the GLTs, an RNA that only contains the switch region of a certain isotype. GLTs are necessary during CSR since they expose ssDNA interacting with the switch region DNA and forming RNA:DNA hybrids which are necessary for AID targeting. During CSR, transcription starts in at least two Ig genes. This transcription is promoted by different stimuli targeting the cytokine-inducible promoter (Table 1). Unlike any other Ig isotype, the generation of IGHM GLT is not cytokine-inducible but constantly produced by the enhancer μ (EM). In fact, when the EM is located in another part of the IGH locus, it induces the constant transcription of any downstream gene (*54*). The transcription of the Ig is followed by the splicing of the RNA between the VDJ and the IGH exon. Upon splicing, the spliced-out switch region RNA is closed in DNA lariats. The DNA lariat is opened by RNA debranching enzyme 1 (DRB1), which linearizes the RNA and forms GLTs. Due to the highly repetitive sequences and the GC-richness, GLT usually forms secondary structures (G-quadruplex) that can also drive AID toward the switch region (*55*) (Figure 3B).

Immunoglobulin gene	**Half-life (days)**	**Cytokine promotion of GLTs**
IGHM	10	Constant GLT production
IGHD	3	(not clear)
IGHG3	7	CD40L+IL4+IL13 or IL6
IGHG1	21	CD40+IL10 or IL6
IGHA1	6	IL4+ TGFβ
IGHG2	20	IL6

IGHG4	21	CD40L+IL4+IL13 or IL6
IGHE	2	CD40L+IL4+IL13
IGHA2	6	IL4+ TGFβ

Table 1. Immunoglobulin effector functions and cytokines that trigger the GLT production *(*1*,* 54*,* 56–60*)*.

AID is an enzyme that deaminates cytosine (C) and converts it to uracil (U). AID acts within the WYCR (Adenine/Thymine + Cytosine/Thymine + Cytosine + Adenine/Guanine) motif, found in the switch regions and other parts of the genome; hence, AID also has off-target activity (*61*, *62*). Once AID deaminates the C and becomes U, a mismatch is generated since U cannot match the G in the complementary DNA strand (Figure *3*B). The mismatch is rapidly processed by base excision repair (BER), in which uracil-DNA glycosylase (UNG) removes the U from the DNA and creates an abasic site. The abasic site is nicked by apurinic-apyrimidinic endonuclease 1 (APE1), sometimes in combination with exonuclease 1 (EXO1), creating a blunt dsDNA break that cNHEJ will repair. This abasic site can also be processed by mismatch DNA repair protein type A (MutSa), which would follow the DNA break repair by the alternative end-joining (aEJ) (*63*) (section 4.5.2). AID targeting occurs in two distant switch regions in parallel, generating two dsDNA breaks to be repaired. cNHEJ or aEJ would primarily ligate the two distant dsDNA breaks once they are brought up close together, facilitated by the chromatin loop extrusion (*64*) (Figure *3*B).

The chromatin loop extrusion was first discussed for V(D)J recombination and theorized later for CSR (*64–66*). The chromatin loop extrusion is mediated by cohesin that modulates the extrusion of chromatin between two CCCTC-binding factors (CTCF) sites and generates loop-shaped contact domains. The CTCF sites in the IGH are located in the EM and the 3'regulatory region (3'RR) (*67*). In agreement with different studies using chromatin interaction experiments, it was shown that the EM and 3'RR are in contact in non-activated B cells. Interestingly, the transcription of these two elements is also active at that stage (*64*, *68*). Upon B cell activation and, thus, cytokine induction, the transcription of another switch region starts, and interactions between the EM, 3´RR, and the newly transcribed switch region occur. While the GLT production is ongoing within the chromatin loops, AID would mediate the dsDNA breaks in SM and the cytokine-induced switch region driven by the GLT G-quadruplex in the exposed ssDNA. Cohesin-mediated loop extrusion aligns the dsDNA breaks in the switch regions for deletional end joining, preferably by cNHEJ or aEJ (section 4.5) (*64*, *69*). The dsDNA break resolution gives rise to a switch-joint and the deletion of the DNA between the SM and the other switch region. The switch-joint is formed by parts of SM joined to another switch region. The DNA degradation between both switch regions is encapsulated into a DNA switch circle that decays rapidly when activation stimuli are withdrawn from the B cells (*70*, *71*) (Figure *3*B).

CSR is the diversification process widely accepted to happen during the GC reaction in the DZ. However, it has been observed that one day after Tfh cell

activation and before the GC is anatomically formed, GLTs peaked in transcription, followed by AID expression and induction of CSR (*72*). Also, CSR is not a process that occurs only once in the B cells; indeed, CSR can happen more than once within the same allele of a B cell in a process called sequential switching. Sequential switching was first described in the canonical CSR to IgE, which involves first switching from SM to switch γ1 region (SG1), and secondly between SM-SG1 to switch ε region (SE) (*68*, *73*). Even though sequential switching has not been deeply studied, it has been observed to happen not only to class switch to IgE but also to class switch to IgGs (*74*) and IgAs (*75*).

CSR is a complex process often used as a readout of DNA repair. CH12-F3 (CH12) is a murine lymphoma B cell line used in DNA repair research due to being a B cell line able to carry CSR (*76*). The role of proteins such as P53 Binding Protein 1 (53BP1) and Ligase 4 in the orientation-specific DNA repair during CSR (*77*, *78*) or the role of chromatin binding factor in the CSR were found using CH12 (*79*).

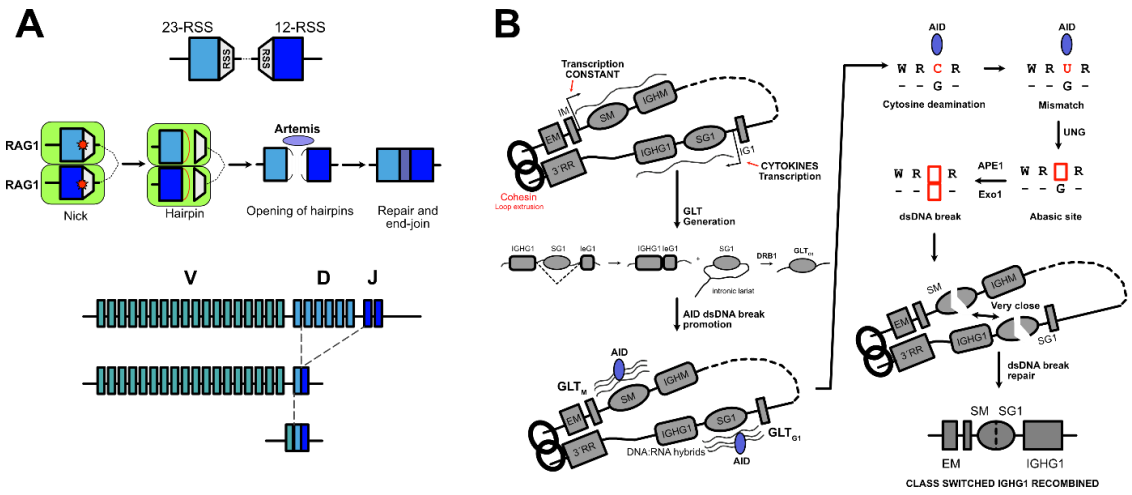


Figure 3. Antibody diversification.

A. V(D)J recombination. Top, depiction of domains containing 12-RSS and 23-RSS. In blue are the sequences carrying RSS that belong to the V, D or J domain. Green corresponds to the protein complex formed by RAG1, RAG2, HMGB1 and HMGB2. Artemis is a cNHEJ protein that can open DNA hairpins upon phosphorylation of DNA-PKcs. Bottom, a simplified way of the two steps of VDJ recombination**.** RSS=Recombination signal sequence, RAG1=Recombination activating gene 1, RAG2=Recombination activating gene 2, HMGB1=High mobility group B 1, HMGB2=High mobility group B 2, cNHEJ=classical non-homologous end joining. **B**. Class-switch recombination. Following the arrows, the process goes from top to bottom and left to right. The example of CSR occurs between SM and SG1. Firstly, the EM and 3´RR are the anchor points of a loop the cohesin holds. Upon cytokine induction, GLTs are formed. GLTs drive AID to target the switch region DNA. AID function is deaminating the C into an U that will become a dsDNA break after the BER process. Eventually, two distant dsDNA breaks, in SM and SG1, will be brought

up together, and dsDNA repair machinery will join them to form a switch-joint. EM=Enhancer μ, 3´RR=3´ regulatory region, SM=Switch μ region, IGHM=Immunoglobulin heavy chain μ, SG1=Switch γ1 region, IGHG1=Immunoglobulin heavy chain γ1, GLT=Germline transcripts, AID=Activation-induced cytidine deaminase, W=Adenine or Thymine, R=Adenine or Guanine, C=Cytosine, G=Guanine, U=Uracil, Y=Cytosine or Thymine, BER=base-excision repair, UNG=Uracil-DNA glycosylase, APE1=Apurinic-apyrimidinic endonuclease 1, EXO1=Exonuclease 1, IG1=Promoter of IGHG1, IM=Promoter of IGHM.

4.3.3. Somatic hypermutation

SHM is the process by which AID deaminates cytosines in the V(D)J recombined domain and generates single nucleotide mutations in an attempt to acquire higher affinity towards an antigen during GC reaction in the DZ. AID is recruited to the V(D)J domain by the motif W**GC**W, especially in areas where several of these motifs are concentrated as a hotspot. AID deaminates the C, in the same fashion as during CSR, creating an U and a mismatch that needs to be resolved. Resolution is processed by non-canonical BER (nc-BER) or non-canonical mismatch repair (nc-MMR) (*80*, *81*).

Even though SHM and CSR are primed by AID activity, different environmental components separate the destiny of the U:G mismatches. It was described that the C-terminus of AID is responsible for CSR, while the N-terminus of AID is responsible for SHM (*38*). Also, APE1 protein from BER is only necessary for CSR but not SHM (*82*). Interestingly, WGCW hotspots also recruit DNA polymerase (Pol) η, a low-fidelity polymerase that fills gaps during nc-BER or nc-MMR repair (*83*). To date, it is unclear how CSR and SHM are entirely regulated.

4.3.4. Regulation of the diversification processes

The diversification processes of the antibody require the generation of mutations to gain affinity towards antigens. However, mutations are also hazardous since they can cause genome stability or generate autoreactive antibodies. Therefore, antibody diversification is tightly regulated by processes to avoid the generation of malignant and self-destructive B cells.

Firstly, distinct mechanisms divided into peripheral and central tolerance ensure the deletion or inactivation of B cells that are reactive to self-antigens, also called autoreactive B cells. The V(D)J recombination is a random deletional recombination process that can cause highly autoreactive antibodies. B cells can be reactive to self-antigens with high or weak affinity. When the affinity towards a self-antigen is weak, the B cell egresses the BM in an anergic form; in other words, it will not be able to respond to antigen encounters. B cells that egress the BM are supervised by peripheral tolerance, which regulates these anergic B cells and leads to apoptosis upon lack of positive selection. When B cells have a high affinity towards self-antigens in the early stages of the B cell development, the central tolerance takes care of reconducting or eliminating the B cell. The central tolerance directs a highly affine autoreactive B cell towards apoptosis or receptor

editing. The apoptosis occurs mainly in the newly formed B cells and is driven by the expression of B cell lymphoma 2–interacting mediator (BIM). However, if B cell lymphoma 2 (BCL-2) is expressed, the pro-apoptotic function of BIM is inhibited, and the receptor editing direction is followed. The receptor editing pathway involves the upregulation of RAG1 and RAG2 and the downregulation of the BCR. The receptor edition is done in the IGL in both alleles, making possible the co-expression of two different IGLs that may lead to the escape of partially autoreactive B cells from the BM. Around 20% of the B cells in BM are estimated to undergo receptor editing (*84*). It is worth mentioning that receptor editing has also been reported to happen during peripheral tolerance (*85*). Receptor editing is possible because the recombination of the IGH and IGL is done only in one allele at a time, thanks to the process of allelic exclusion.

The allelic exclusion guides the monoallelic recombination of the BCR. In other words, the three antibody diversification processes only occur in one allele at a time. It is unclear if CSR and SHM are subjected to allelic exclusion. Studies in cancerous cells or *in vitro* show that CSR and SHM are not subjected to allelic exclusion (*86*, *87*). Although transcription is found in both IGH alleles, further molecular players are necessary to define if CSR occurs upon IGH transcription (*88*).

Epigenetics tightly regulate allelic exclusion. Epigenetically, the recombined allele is hypomethylated, while the excluded allele is hypermethylated, impeding the access of proteins to it. Nevertheless, the transcription of both alleles has been found to happen simultaneously in the B cells independently of the allelic exclusion process (*89*). If the recombination of an allele gives rise to a successfully expressed antibody, further rearrangements are prevented by the downregulation of RAG1, RAG2 and TdT (*90*). However, further rearrangements within the same allele or the other allele would follow if it fails (*91*). Allelic exclusion confers two opportunities for a B cell to rearrange its VDJ into a successful antibody. Allelic exclusion can also be considered an early cell-autonomous tolerance mechanism that helps reduce the chance of B cells expressing two poly-specific BCRs, facilitating central tolerance surveillance (*92*).

4.4. Inserts in the IGH locus

The dogma of antibody diversification comprises the three processes described in the previous section: V(D)J recombination, CSR and SHM. However, a new mechanism of antibody diversification was discovered in 2016 in which the IGH of the antibody acquires DNA fragments in the switch regions from all over the genome (*93*). Tan et al., 2016 discovered that inserts could confer a broader reactivity against different subtypes of malaria parasites. Leucocyte-associated immunoglobulin-like receptor 1 (LAIR1) was found in the antibodies of up to 10% of malaria-exposed donors. The genomic insert comprised the LAIR1 exon flanked by its introns enabling splicing into the antibody transcript and, thus, translated (*94*). Further research in malaria-exposed donors also discovered leucocyte immunoglobulin-like receptor B1 (LILRB1)-containing antibodies of

several malaria-exposed donors (*95*). Interestingly, LAIR1- and LILRB1-containing antibodies provided a broadly reactive immune response against several variants of the malaria parasite, in contrast to their non-insert antibody counterparts. These insert-containing antibodies were broadly reactive because repetitive interspersed families of polypeptides (RIFIN), malaria parasite surface antigen, binds to LAIR1 and LILRB1 to activate immune suppression and therefore, the antibody acted as a decoy attracting the parasite to them (*96*). Thus, the antibody would have acquired RIFIN-interacting proteins domains and increased the chance of encountering the parasite. Even though the mechanism remains unknown, it was suggested to be a copy-paste mechanism since, in the case of the LAIR1 insert, the original LAIR1 gene was found intact in both alleles (*93*). Further research in the mechanism of insert incorporation must be done to bring light to this new B cell diversification mechanism.

These inserts were found between (-): V-DJ, VD-J, J-CH1, V-J, V-CH1 and V-D-J (*97*). Inserts within the VDJ segments strongly originated from introns, while J-CH1 inserts originated from exons. Mitochondria original inserts were found only between VDJ segments and originated mainly from the D-loop part of the mitochondrial genome (*97*). The inserts do not always translate to the antibody protein due to the splicing process or stop codons. The splicing removes the inserts incorporated in intronic regions, such as between J-CH1 segments. Inserts originating from distant genes were detected in the IGH of European donors in genomic DNA and RNA. The insert origin correlated with the transcriptional landscape of B cells, suggesting a link between inserts and transcription (*94*, *97*). Even though insert frequency is very low, they contribute to the diversification of antibodies and should be considered part of the B cell diversification process.

4.5. DNA repair mechanisms of double-stranded DNA breaks

The genome acquires breaks regularly due to endogenous processes, e.g., DNA replication, and exogenous processes, e.g., ultraviolet (UV)-light. Erroneously repaired DNA breaks result in mutations, namely DNA nucleotide substitutions, small or large insertions or deletions. Cells have evolved distinct DNA repair mechanisms to avoid the damaging outcomes of undesired mutations and unrepaired broken DNA.

The first response to dsDNA break is to recruit ataxia-telangiectasia mutated kinase (ATM). ATM attracts the Mre11-RAD50-NBN (MRN)-complex (*98*). The MRN-complex is an anchor between the two sides of the DNA break in cNHEJ and aEJ. It also bridges between the DNA break and the sister chromatid in homologous recombination (HR) (*99*). ATM phosphorylates histone family member X (H2AX) to become γH2AX. γH2AX acts as a positive feedback loop recruiting more ATM to phosphorylate more H2AX. Accumulation of γH2AX in the DNA lesion localize the DNA repair action in a "foci" (*100*).

The dsDNA break has different resolutions depending on the type of break and the cell phase when it is produced. The different pathways could be divided into error-prone and error-free repair.

4.5.1. Classical non-homologous end joining

cNHEJ is an error-prone dsDNA break repair mechanism. It is known for its ability to join ends without needing homology (blunt ends) or using less than 3 nt of homology. cNHEJ is the fastest dsDNA break repair mechanism and the most used by mammalian cells since it is active throughout the G1, S and G2 phases (*101*, *102*).

In cNHEJ, γH2AX recruits 53BP1, which gets phosphorylated by ATM in Serine 25 (*103*). Replication timing regulatory factor 1 (Rif1) translocates to the break in a 53BP1 phosphorylated-dependent way. Rif1 is a protein that protects the DNA breaks from resection, which contributes to the standard processing of the cNHEJ and inhibits HR (*104*, *105*). In parallel, chromatin remodeling occurs: Histone 3 (H3) and 4 (H4) get acetylated and recruit Ku70/Ku80 heterodimer, which protects the dsDNA break ends (*106*). The Ku70/Ku80 heterodimer has a ring-like form that enables the heterodimer to hug the DNA break ends dynamically. DNA-PKcs are recruited and interact with Ku70/Ku80, activating its Serine/Threonine kinase. Physically there are two DNA-PKcs kinases, one at each DNA break. DNA-PKcs get phosphorylated, either through autophosphorylation or by ATM. The two phosphorylated DNA-PKcs form a synapsis, making the two dsDNA breaks come closer together. In some cases, the DNA ends are ligated to the complementary strand, forming a hairpin. Resolution of the hairpin starts with DNA-PKcs phosphorylation of Artemis, which will subsequentially cleave the hairpin. If the DNA end has a 3'overhangs, there would be a nucleotide gap-filling process to fill the missing strand by DNA Pol η or DNA Pol μ. These two DNA polymerases are highly imprecise, resulting in inserts and deletions (indels) (*107*, *108*). Blunt ends are captured by the complex formed by X-ray repair cross-complementing protein 4 (XRCC4) and Ligase 4, and lastly, get ligated (*109*) (Figure *4*).

4.5.2. Alternative end-joining

The aEJ, also known as microhomology-mediated end joining (MMEJ), is an error-prone DNA repair pathway that shares the end resection with HR and the non-homologous joining with cNHEJ. It is a DNA repair that is not highly used by mammalian cells, but it is of high importance because its usage promotes genomic instability by, for example, translocations. aEJ is functional throughout the G1, G2 and S phases (*110*, *111*).

In the first step, poly [ADP-ribose] polymerase 1 (PARP1) and 53BP1 are recruited to the dsDNA break. Then, CtBP-interacting protein (CtIP) in combination with meiotic recombination 11 (Mre11), are recruited to the 3´overhangs where the process of resection starts (*110*). Resection comprises the degradation of nucleotides from a blunt end to produce an ssDNA or overhang. Thus, Mre11 resects 3'single-stranded regions using the 3'-to-5'-

exonuclease activity to expose microhomology (MH). Replication protein A (RPA) binds the exposed ssDNA. Rapidly, DNA repair protein RAD52 homolog 2 (RAD52) and DNA Pol θ bind the ssDNA and promote the annealing of complementary ssDNA strands (*112*, *113*). As soon as MH is found, Mre11 pauses the resection for a proper alignment of sequences. The MH can comprise an alignment of 3 to 20 nt. The MH exposes the non-homologous 3'-tails that excision repair cross complementation group 1 (ERCC1) and xeroderma pigmentosum group F-complementing protein (XPF) nucleases digest (*114*). Resulting gaps within the DNA strands are filled by DNA Pol 4, DNA Pol μ, DNA Pol η and DNA Pol ζ. DNA ends are finally ligated by the complex formed by X-ray repair cross-complementing protein 1 (XRCC1) and Ligase IIIa or Ligase I (*111*, *115*) (Figure *4*).

aEJ has been shown to be the second most preferred mechanism used during CSR after cNHEJ. The second most frequent type of CSR junctions are those bearing 1-4 nt microhomologies, which are the typical genomic scars found after aEJ break repair (*69*).

4.5.3. Homologous recombination

HR is an error-free or high-fidelity DNA repair mechanism that repairs DNA breaks occurring during S and G2 phases. It uses the sister chromatid as a template and is the slowest DNA repair pathway (*69*).

Mre11, as part of the MRN complex, resects up to 300 nt enhanced by CtIP, similar to aEJ. Breast cancer gene 1 (BRCA1) in complex with BRCA1-associated RING domain protein 1 (BARD1) acts as an enhancer of Mre11-CtIP. The 5'-tail formed is an entry point for endonucleases like EXO1, DNA Replication ATP-Dependent Helicase/Nuclease DNA2 (DNA2) or 3′−5′ ATP-dependent RecQ DNA helicase (BLM) that make the break have a long 3'-tail. RPA covers and protects ssDNA from the 3´-tail from degradation. Breast cancer 2 (BRCA2) binds the ssDNA, displaces RPA, and starts binding to BRCA1, BARD1 and recruiting partner and localizer of BRCA2 (PALB2). DNA repair protein RAD51 homolog 1 (RAD51) is recruited to the DNA break by the DNA enrichment of BRCA1, BRCA2 and PALB2. However, it can also be loaded into the DNA in a BRCA2-independent way. The BRCA2-independent RAD51 loading is regulated by polo-like kinase 1 (PLK1) and casein kinase 2 (CK2). They both phosphorylate RAD51 in Serine14 and Threonine13 ($RAD51^{S14T13}$). $RAD51^{S14T13}$ binds to the MRN-complex right after DNA break sensing. Even though RAD51 is constitutively expressed throughout the cell cycle (*116*), phosphorylated $RAD51^{S14T13}$ peaks at the end of G2, which makes it accumulate in DNA breaks and form foci during G2 and S phases (*117*). RAD51, physically and not enzymatically, promotes HR and inhibits any other DNA repair (*118*). RAD51-ssDNA is a dynamic nucleoprotein able to screen for homology in sister chromatids. When the homology is found, a three-strand synapsis stabilizes, and DNA Pol δ synthesizes a new DNA strand using the sister chromatid as a template (*119*) (Figure *4*).

HR is considered dispensable for CSR; however, deletion of crucial HR molecular players, e.g., RAD51, causes the persistence of breaks in the IGH in post-

replicative B cells. It has been hypothesized that HR rescues IGH genes whose breaks were not efficiently repaired by cNHEJ. Considering that HR would use the sister chromatid to repair the IGH, HR would essentially rescue the IGH with the possibility of rearranging it once again (*69*).

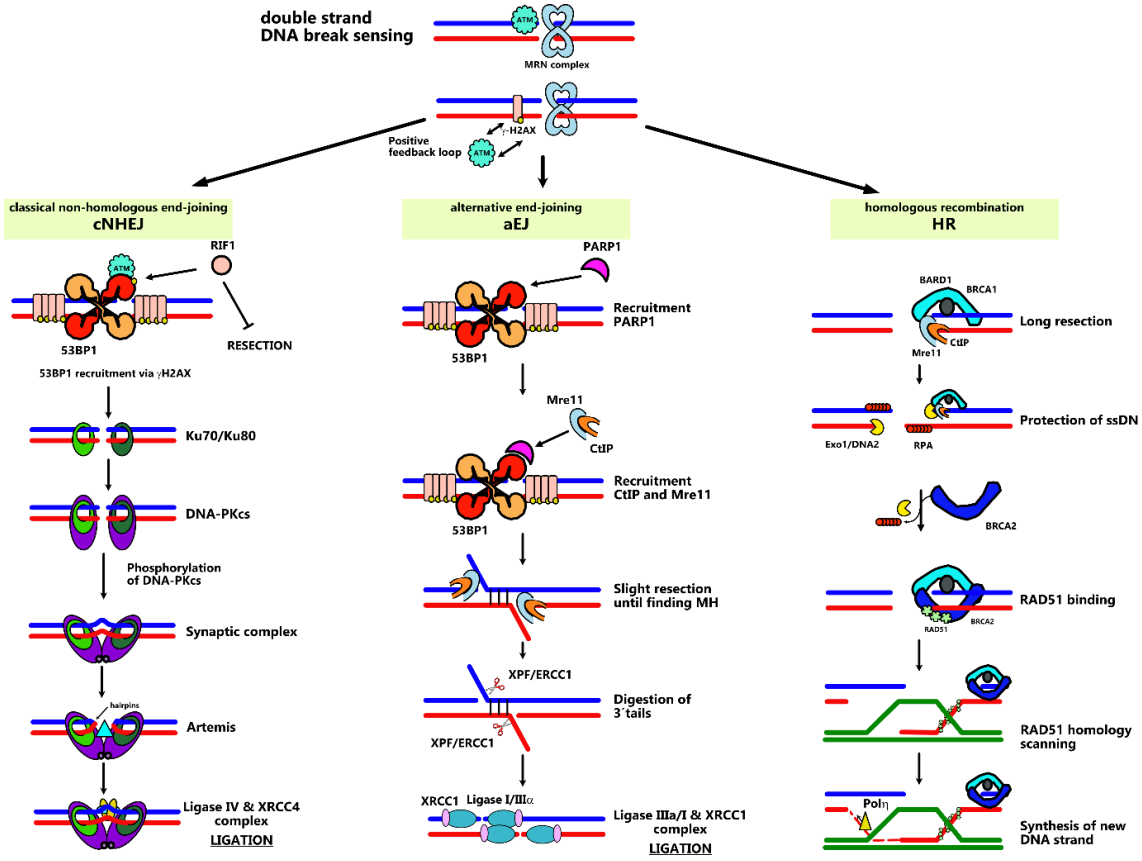


Figure 4. dsDNA break repair mechanisms

At the top is a representation of the double-strand DNA break-sensing process. From left to right, we find cNHEJ, aEJ and HR DNA repair processes. 53BP1=P53 Binding Protein 1, Rif1=Replication timing regulatory factor 1 , DNA-PKcs=DNA-dependent protein kinase, catalytic subunit, XRCC4=X-ray repair cross-complementing protein 4, ATM=Ataxia-telangiectasia mutated kinase, MRN=Mre11-RAD50-NBN, H2AX=Histone family member X , PARP1=Poly [ADP-ribose] polymerase 1, Mre11=Meiotic Recombination 11 , CtIP=CtBP-interacting protein, MH=Microhomology, XPF=Xeroderma Pigmentosum Group F-Complementing Protein, ERCC1=Excision repair cross complementation group 1, XRCC1=X-ray repair cross-complementing protein 1, BRCA1=breast cancer gene 1, BRCA2=breast cancer gene 2, BARD1=BRCA1-associated RING domain protein 1, RPA=Replication protein A , ssDNA=single-stranded DNA, EXO1=Exonuclease 1, DNA2=DNA Replication Helicase/Nuclease 2 , RAD51=DNA repair protein RAD51 homolog 1, Pol=DNA polymerase.

4.6. Telomerase and telomere elongation

DNA repair maintains the cell alive by correctly repairing the DNA breaks. However, a cell has a limited number of cell cycles before it enters apoptosis, and that is due to the length of the telomeres. In every DNA replication, the telomeres shorten; thus, there is a point where there are no more telomeres, and the coding genome starts to be shortened (*120*, *121*). In fact, long telomeres are essential for highly proliferative cells such as GC B cells. GC B cells are expected to undergo several cycles of proliferation to maturate their antibody and one of the aspects that intrinsically keeps them alive is the telomeres. Elongation of telomeres is done via the telomerase reverse transcriptase (TERT). Naïve B cells express high TERT before entering the GC reaction, which has been accounted to the fact that GC B cells proliferate extensively during the GC reaction; therefore, elongating the telomeres ensures genome stability during this process (*122*, *123*).

The telomeres are a region of repetitive sequences (TTAGGG) found at the ends of each chromosome. TERT is an RNA reverse transcriptase that works with the human telomerase RNA (hTR) to elongate telomeres. hTR (CAAUCCCAAUC) forms secondary structures and couples with TERT to serve as a template for telomere elongation. The elongation of telomeres is highly regulated by TERT expression, telomere capping (Shelterin) and telomere extension abrogation (CTS Telomere Maintenance Complex Component 1(CTC1)-Oligosaccharide Binding Fold Containing 1(STN1)-Telomere Length Regulation Protein 1 (TEN1) (CST complex)) (*124*).

TERT is highly regulated and is expressed in progenitor cells, such as stem cells or BM progenitors. Also, it has been observed that renewing tissues express TERT, while cells like fibroblasts do not express it, entering a state of senescence (*124*). On the other hand, Also, cancer cells are known to reactivate TERT to have an unlimited telomere length (*125*).

The telomere ends are capped and protected by a protein complex, Shelterin, to prevent the naked ends from activating the DNA damage response. Tripeptidyl-peptidase 1 (TPP1), a protein in Shelterin, recruits TERT to the telomeres. TPP1 is bound to the protection of telomeres protein 1 (POT1), which binds the ssDNA of the telomeres. It is thought that hTR and POT1 compete for ssDNA binding and regulates TERT activity. Only when hTR is bound to the ssDNA would TERT start the process of telomere elongation (*126*).

The telomere elongation follows a process called "repeat addition processivity." hTR contains two telomeric repeats. Telomerase initially binds its telomere DNA substrate via base-pairing interactions to the first telomeric repeat of hTR. Once bound, a single telomere DNA repeat is rapidly synthesized. The newly formed nucleotides align with the first repeat of hTR, giving space to synthesize a new one (*127*). The process continues to elongate until the inhibitory CST complex is recruited together with DNA Pol α. The CST complex sequesters the single-stranded telomeric overhang and promotes the DNA synthesis of the complementary strand by DNA Pol α (*128*).

TERT has off-target activity outside the telomeres that have been observed as interstitial telomeric sequence (ITS). ITS are telomeric repeats included at genomic scars after dsDNA break repair all around the genome. dsDNA repair proteins can be found in telomeres, and telomeric repeat factor 2 (TRF2), a protein part of Shelterin, is often recruited by the DNA repair machinery. ITS is often flanked by 1-5 nt of homology to the hTR, which suggests the involvement of TERT in their synthesis (*129*, *130*). On the other hand, TERT can also be found in mitochondria. Human TERT contains a mitochondrial signal peptide sufficient for mitochondrial localization (*131*). TERT activity in mitochondria has been linked to the protection of mitochondrial DNA (mtDNA) and the downregulation of reactive oxygen species (ROS) (*132*).

Altogether, TERT is the human RNA reverse transcriptase that maintains the telomeres and protects them from shortening, but at the same time, it has non-canonical functions as part of the ROS system regulation.

4.7. Cancer and DNA repair mechanism

The genome is constantly exposed to damage from exogenous and endogenous sources, such as UV-light or antibody diversification processes in B cells. Unrepaired DNA damage can direct the cell fate to apoptosis and mutations driving carcinogenesis (*133*). Healthy cells roughly bear one mutation per 100 billion bases in their genome; however, if they suffer DNA repair malfunction, the mutation rate increases together with the chances of developing disease (*134*). For instance, inborn errors in Ligase 4 cause a B cell immunodeficiency since the normal antibody diversification processes are impaired due to the inefficient repair of DNA breaks (*135*, *136*).

DNA repair malfunction is commonly caused by single nucleotide polymorphisms (SNP) or deletions in genes of proteins that are part of the DNA repair machinery. It promotes the accumulation of DNA lesions, driving the malignant transformation of cells. Indeed, cancer cells are estimated to carry 3.16 mutations per megabase. Mutations can be classified as "driver" or "passenger" mutations. "Driver" mutations provide an advantage to somatic cells in their microenvironment, thus driving them to cancer, while "passenger" mutations are defined as mutations that do not provide any proliferative benefit. Controversially, most mutations found in cancer cells are "passenger" mutations (*137*, *138*).

Mutations in cancer cells are crucial to research for characterizing tumors and choosing the right therapy to combat them. There are two types of chemotherapies: i) targeting and ii) non-targeting. Targeted chemotherapies were developed toward specific molecular targets on cancer cells, while non-targeted chemotherapies generally drive the death of highly proliferative cells (*139*). Unlike targeted, non-targeted chemotherapies, like platinum or radiation, are cytotoxic and drive mutation acquisition in healthy cells as well. Considering that cancer patients likely have DNA repair malfunctions, it is counterproductive to treat them using DNA break-inducing chemicals that would create mutations in non-cancerous cells, potentially driving malignancies in other tissues (*140*).

Cancer prognosis is most commonly assayed in clinics using SNP screening. SNP screening is an assay based on sequencing the most common SNPs found in people who developed cancer, e.g., BRCA1 mutation. An improved cancer prognosis assay relates somatic cells' mutational rate to known mutational landscapes of cancer cells, identifying potential somatic cells that will develop cancer (*141*).

Cancer development and DNA repair malfunction are tightly linked. All in all, cancer prognosis is commonly done by screening DNA repair protein mutations. Every year, cancer research discovers new mutations that initiate or maintain cancer linked to DNA repair proteins, challenging screening relevant DNA repair proteins. The future holds for analyzing DNA repair malfunction without screening all DNA repair proteins individually, most optimally studying DNA repair genomic scars.

5. AIMS

The overall aim of my Ph.D. thesis was to i) develop a methodology to study CSR genomic scars and apply it to ii) study the insert incorporation mechanism and iii) the identification of DNA repair malfunction.

5.1. Development of a methodology to characterize B cell clones, genomic inserts, and DNA repair outcomes by analyzing CSR genomic scars

The current methodologies employed for studying CSR junctions are significantly limited in their ability to characterize and analyze the complexity exhibited by these junctions accurately. On the one hand, the length limitation of short-sequencing does not allow the analysis of complex switch-joints, such as sequential switching or inserts-containing junctions, since their length can span the maximum length of high-throughput sequencing. On the other hand, these analyses are often standardized using the total number of reads. However, this approach is not representative of the B cell diversity because i) PCR amplification favors the amplification of smaller amplicons resulting in an overrepresentation of small clones, and ii) the same B cell clones can be overrepresented due to illness, vaccination or endogenous immune responses in a polyclonal B cell sample.

Therefore, the first aim of my thesis was to develop a methodology to analyze switch-joints establishing a PCR using the minimum number of primers and long-read sequencing. Unlike current methods, the novel method would involve the identification of inserts and unique B cell clones.

5.2. Addressing the insert incorporation mechanism in the antibody heavy chain locus

Antibody inserts have emerged as a novel area of research. Genomic inserts in the antibody gene are rare but potentially valuable as they can broaden the affinity of antibodies. Despite being thoroughly studied in antibody transcripts, the mechanism of insert incorporation remains unknown. Therefore, in my thesis, I aimed to identify crucial DNA repair proteins involved in the acquisition of inserts in switch-joints in culture and patients. *in vitro* and *in vivo*.

To investigate insert acquisition in antibody genes, I focused on i) insert source and ii) the DNA repair mechanism involved. Originally, inserts were found in the proximity of ERFS. To investigate the involvement of genomic instability, I promoted ERFS breakage using hydroxyurea. Moreover, I targeted the two preferred mechanisms in CSR to study the DNA repair mechanism involved in insert ligation. I affected the cNHEJ inhibiting Ligase 4 and aEJ inhibiting PARP1, and in general, I affected dsDNA break sensing inhibiting ATM. In addition and assuming that telomerase can off-target switch regions in the same fashion as AID off-target telomeres, I analyzed the role of telomerase in the insert acquisition by inhibiting it *in vitro*. Bearing that inhibition of proteins *in vitro* does not equal partial or total absence of protein *in vivo*, I analyzed DNA repair protein-deficient patients to observe the natural insert acquisition in the absence of ATM, Ligase 4, nipped-B-like protein (NIPBL), AID and BRCA1.

5.3. Profiling breakpoints and analyzing DNA repair efficiency in CSR genomic scars of human and mouse samples

DNA repair malfunction is routinely diagnosed by screening target polymorphisms in DNA repair proteins linked to the initiation or maintenance of cancerogenesis, such as BRCA1 and BRCA2. Such screenings are hampered by the fact that our knowledge of the molecular players involved and the polymorphisms is limited. Thus, I aim to study DNA repair malfunction by analyzing the outcomes of DNA repair using switch-joints as somatic genomic scars in humans and mice. Considering that all main DNA repair pathways are active during CSR, this analysis would reflect how DNA breaks are repaired in the body.

The final aim of this thesis was to separate healthy from DNA repair-deficient samples using the analysis of switch-joints using either a i) visualization or ii) sequence analysis.

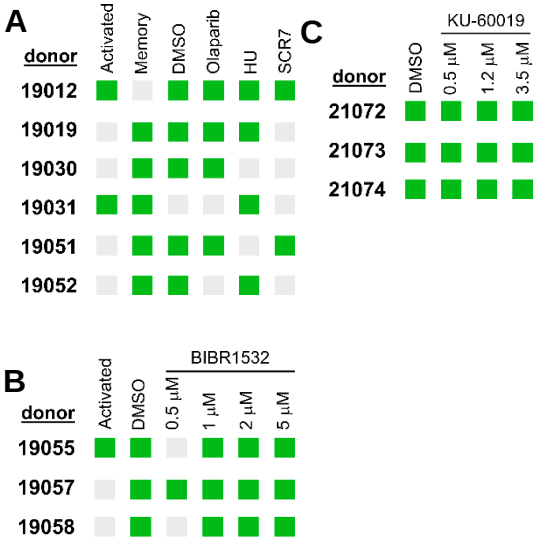
6. MATERIAL AND METHODS

6.1. B cell culture

6.1.1. Primary human B cells

Primary human B cells were cultured using 10% fetal bovine serum (FBS) (PAN Biotech GMBH, #P30-3302), 1% non-essential aminoacids (Life Technologies, #11140035), 1% sodium pyruvate (Life Technologies, # 11360039), 1% GlutaMAX (Life Technologies, #35050038), 1% Penicilin-Streptomycin (Life Technologies, #15140163), 0.1% mercaptoethanol (Life Technologies, #31350010), 5 µg/ml Kanamycin (Serva Electrophoresis, #26897.01) and 0.002% Transferrin (Merck-Milipore, #616397) in 1M 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) Roswell Park memorial institute medium (RPMI) (Life Technologies, #42401018).

Primary human B cells were cultured in 12-wells plates (SARSTEDT, #833921) at a concentration of 200 000 cells per ml with four grades (Gy)-irradiated CD40L expressing K562L cells to induce their activation. The culture was carried out for seven days. On days 0, 1, 3, and 5, 25 nanograms (ng) of IL4 (recombinantly produced) was added to each well to trigger B cell activation. If the primary human B cells were treated with some chemicals (Figure 5C), they were to the well added simultaneously with the IL4. Cells were harvested on day 7. gDNA isolation of primary human B cells was performed using the ReliaPrep™ Blood genomic DNA (gDNA) Miniprep System (Promega, #A5082).



D

Chemical	Company, catalog (#)	Concentration used
DMSO	Roth #A994.1	Max DMSO used in the experiment
Olaparib	BIOZOL #FBM-10-2154	1.25 µM
Hydroxyurea (HU)	Th.Geyer Berlin #118142008	25 µM
SCR7	Hölzel Diagnostika #A15447-5	80 µM
BIBR1532	Hölzel Diagnostika #M1676	0.5, 1, 2, 5 µM
KU-60019	Hölzel Diagnostika #TMO-T2474	0.5, 1.2, 3.5 µM

Figure 5. Description of experiments done with chemicals.

A. Chemicals used in each donor belonging to the "DNA repair modulation" experiment. HU=Hydroxyurea. µM=micromolar. **B.** Chemicals used in each donor belonging to the "telomerase inhibition" experiment. **C.** Chemicals used

in each donor belonging to the "ATM inhibition" experiment. **D.** Chemicals used in experiments. In the left column is the chemical's name; the middle column, the company and catalog number and the right column used concentrations in the experiments.

6.1.2. CH12 cells

CH12 cells were cultured using 10% FBS (PAN Biotech GMBH, #P30-3302), 1% non-essential aminoacids (Life Technologies, #11140035), 1% sodium pyruvate (Life Technologies, # 11360039), 1% GlutaMAX (Life Technologies, #35050038), 1% Penicilin-Streptomycin (Life Technologies, #15140163), 0.1% mercaptoethanol (Life Technologies, #31350010), 5 μg/ml Kanamycin (Serva Electrophoresis, #26897.01) and 0.002% Transferrin (Merck-Milipore, #616397) in 1M HEPES RPMI ((Life Technologies, #42401018).

CH12 cells were cultured in 12-well plates (SARSTEDT, #833921) at a concentration of 200 000 cells per ml. The culture was carried out for three days. On day 1, 25 ng IL4 (recombinantly produced), 4300 ng of anti-mouse CD40 (BIOZOL, #BLD-102902) and 8.6 ng of mTGF-b1 (R&D Systems, #7666-MB-005/CF) were added to each well to trigger B cell activation. Cells were harvested on day 3. gDNA isolation of CH12 was performed using the ReliaPrep™ Blood gDNA Miniprep System (Promega, #A5082).

The CH12 cell lines were kindly provided by the group of Michela di Virgilio (Table *2*).

CH12 mutation	**Creator**
Ligase $4^{-/-}$ (partial deletion of exon 2)	Anna Guzman
$BRCA1^{-/-}$ (partial deletion of exon 11)	Sandhya Balasubramanian
$53BP1^{-/-}$ (partial deletion of exon 18)	Devakumar Sundaravinayagam
$Rif1^{-/-}$ (partial deletion of exon 2)	Matteo Andreani

Table 2. CH12 mutants used for the analysis of class-switch recombination breakpoint profiles.

6.1.3. DNA repair deficient patients

The human samples of DNA repair deficient patients were supplied by Qiang Pan-Hammerström (Karolinska Institute, Sweden). The exact mutations found in the human samples are collected in Table 3.

Diagnosis	**Mutation**
Ligase $4^{-/-}$	Undefined
$BRCA1^{-/-}$	3171ins5
$NIPBL^{-/-}$	c.6647A>G, p.Y2216C
$ATM^{-/-}$	p.Gln1852ProfsX5
$ATM^{-/-}$	Hom Large deletion
$AID^{-/-}$	Hom p.Arg98X

Table 3. Genotypes of human DNA repair deficient patients.

6.2. Flow cytometry & fluorescence-activated cell sorting (FACS)

The preparation of cells for flow cytometry and FACS was carried out in the same fashion. The only difference was the number of cells used and the device where the staining was performed in. Flow cytometry staining was done in a 96-well U-bottom plate (SARSTEDT, #833925500) with 25 microlitres (µl) of staining solution and 50 000 to 200 000 cells. FACS staining was performed in a 15 ml falcon tube (Faust, #TPP91015) with 800 µl of staining solution and 15 to 50 million cells.

The staining was carried out by centrifuging the cells at 500g for 5 minutes. Then, cells were resuspended in the staining solution that contained the antibodies chosen for the analysis (Table *4*) and incubated for 7 minutes at 4°C in the dark. Cells were rinsed with eight times more volume of MACS buffer and centrifuged for 5 minutes at 500g. Cells were washed once with a 0.5 ng/ul DAPI solution, rinsed with eight times more MACS buffer and centrifuged for 5 minutes at 500g. Cells for flow cytometry were resuspended in 400 µl of MACS buffer and cells for FACS were resuspended in 500 to 1 000 µl of MACS buffer, depending on the number of stained cells.

Flow cytometry was done in LSRFortessa™ Cell Analyzer (BD Biosciences) and FACS in BD FACSAria™ Fusion Flow Cytometers (BD Biosciences). First, single staining of the colors used was used to calculate compensation. Then 50 000 events were recorded from each sample. Analysis was done using FlowJo (BD Biosciences).

Target	Fluorophore	Info	Dilution	Analysis
CD27	PE	Miltenyi Biotec, #130-114-156, M-T271	32	FACS
IgD	PE-Cy7	Miltenyi Biotec, #130-098-584, IgD26	32	
IgM	AF488	Life Technologies, #A21215	160	
IgG	AF647	Dianova, #109-606-170	80	
IgA	AF647	Dianova, #109-606-011	80	
DAPI	UV	BD Biosciences, #564907	2000	
IgM	AF488	Life Technologies, #A21215	100	Flow cytometry
IgG	AF647	Dianova, #109-606-170	500	
IgA	AF647	Dianova, #109-606-011	500	
DAPI	UV	BD Biosciences, #564907	2000	

Table 4. Antibodies used in flow cytometry and FACS for different analyses.

The staining solution was done using MACS buffer. Alexa Fluor=AF, FACS=Fluorescence-activated cell sorting.

6.3. Generation of artificial switch-joint reads

6.3.1. *In-silico* prototype

The *in silico* prototype was created in R to generate artificial switch-joints. The *in silico* prototype generated random switch-joints using the switch region annotations (Table 8). The script used the following R packages: ggplot2, stringr, ggridges, seqinr, Biostrings, DescTools, phylotools and tidyverse.

Firstly, the switch region sequences were loaded into the script in FASTA format using seqinr. Then a loop to create X number of reads took place to make four different types of reads:

- **Deletion reads**: Switch-joints comprised of SM and another switch region with deletion of random size between 1-1 000 nt and the joint between the end of SM and start of the other switch regions.
  1) SM and a random switch region are joined, forming a switch-joint of two complete switch regions:

```
a<-sample(switch_region, 1)
SX<-switches[a]
joint<-str_c(SM, SX)
```

  2) The size of the deletion that the read would have is randomly chosen between 1 and 100 nt:

```
deletion<-sample(seq(1:1000),1)
```

  3) The switch-joint created is then cut SM in the middle and removing nucleotides equal to the size of the random deletion:

```
joins<-str_c(str_sub(joint, end = round(SM_length/2) -
deletion/2), str_sub(joint, start = round(SM_length/2)
+ 1 + deletion/2))
```

- **Random reads**: Switch-joints with random breakpoints in SM and another switch region that is not SM (SX).

  1) SM and a random switch region are included into a vector:

```
a<-sample(switch_region, 1)
SX<-switches[a]
joint<-str_c(SM, SX)
vector<-c(SM, SX)
```

  2) A loop comprising 2 rounds, one per switch region involved, is made to create the switch-joint:

```
for(i in 1:2){
```

  2.1) The length of the SM and the other switch region are considered. Four different positions are randomly selected considering the length of each switch region involved:

```
length<-str_length(paste(vector[i]))
random<-sort(sample(length,4))
```

2.2) In the first round of the loop, the four randomly chosen positions are used to create two stretches of a read of the SM combining positions 1-2 and 3-4. Thus, creating a deletion between positions 2-3:

```
    if(i==1){
  joins<-
  str_c(str_sub(paste(vector[i]),start=sample[1],
  end=sample[2]),str_sub(paste(vector[i]),start=s
  ample[3],end=sample[4]))   }
```

2.3) In the second round of the loop, the four randomly chosen positions are used to create two stretches of a read of the randomly chosen switch region are generated combining positions 1-2 and 3-4. These stretches are added to the SM stretches. Also, the switch region carries a deletion between positions 2-3:

```
    if(i==2){
    joins<-str_c(joins,str_sub(paste(vector[i]),
start=sample[1],end=sample[2]),str_sub(paste(vect
or[i]), start=sample[3],end=sample[4]))   }
```

- **Consecutive reads**: Switch-joint between SM and two other random switch regions. The breakpoints occurring between the switch regions are random.

1) Two random switch regions, SX and SXX, are randomly selected:

```
a<-sample(switch_region, 1)
b<-sample(switch_region[-(1:a)], 1)
SX<-switches[a]
SXX<-switches[b]
vector<-c(SM,SX,SXX)
```

2) A loop comprising 3 rounds, one per switch region involved, is made to create the switch-joint:

```
for(i in 1:3){
```

2.1) The length of the SM and the other switch regions are considered. Two positions are randomly selected considering the length of each switch region involved:

```
 length<-str_length(paste(vector[i]))
 sample<-sort(sample(length,2))
```

2.2) In the first round of the loop, the two randomly chosen positions are used to create a stretch of SM combining positions 1-2:

```
 if(i==1){
joins<-
str_c(str_sub(paste(vector[i]),start=sample[1],
end=sample[2]))     }
```

2.3) In the second and third rounds of the loop, the two randomly chosen positions are used to create a stretch of SX and SXX, respectively, combining positions 1-2. These stretches are added to the SM stretches:

```
 if(i>1){
joins<-str_c(joins,str_sub(paste(vector[i]),
start=sample[1],end=sample[2]))    }
```

- **Consecutive 2 reads**: Switch-joint between SM and three other random switch regions. The breakpoints occurring between the switch regions are random. The script looks the same as described in the "Consecutive reads", but there is an extra step to choose a third switch region. Thus, there are SM, SX, SXX and SXXX.

6.3.2. Naturficial reads

Naturficial reads were generated by Benedikt Obermayer. The naturficial reads are described as naturally templated computationally generated reads. The reference sequences of 10 000 unique B cell clones from 8 HD were obtained. Samples comprising a different amount of clones and reads were made. The naturficial reads were mutated randomly following the MinION mutations observed in the non-PCR MinION control (Table 5).

	A	C	G	T
A	0.190063	0.000193921	0.0116397	0.000554028
C	0.000574434	0.292938	0.000341318	0.00207396
G	0.0112465	2.17643e-05	0.288772	0.000121508
T	0.001054	0.00121019	8.39998e-05	0.199111

Table 5. Mutation rates between nucleotides to generate naturficial reads.

Mutations were obtained from a sample run in MinION without PCR (the SM plasmid).

These mutations comprised single nucleotide mutations (Table 5) and indels (chance to get an insertion=0.0200868 with an increasing size probability of 0.511982 and a deletion=0.0253724 with an increasing size probability of 0.687357).

6.4. Primary human B cell isolation from blood

Buffy coats were purchased from the DRK-Butspendedienst Nord-Ost GmbH. All buffy coats were free from antibodies against human immunodeficiency virus (HIV), hepatitis C virus (HCV) and hepatitis B virus (HBV).

Firstly, 80 ml of blood was mixed with 20 milliliters (ml) of 2 millimolar (mM) ethylenediaminetetraacetic acid (EDTA) phosphate-buffered saline solution (PBS-T) (Promega, #V4231 & Biotek, #EL406). The blood was distributed in 4 falcons, slowly pouring 30 ml of the blood on top of 15 ml of Ficoll (Carl Roth GmbH, #0642.2). Falcons are centrifuged for 25 minutes (min) at 500 grams (g) with an acceleration/break of 3/0 (Centrifuge Eppendorf 5910R). Blood was separated density-wise and lymphocytes were isolated from the white layer between the plasma and the Ficoll. The isolated lymphocytes were washed twice with 2 mM EDTA PBS-T.

Primary human B cells were isolated using CD19 Microbeads (Miltenyi Biotech, #130-050-301). Therefore, 500 million lymphocytes were incubated with 200 µl CD19 Microbeads in ice for 15 minutes in 800 µl, followed by a washing step using 3 ml MACS buffer (10% FBS 2 mM EDTA PBS-T) and a centrifugation step. Lymphocytes were resuspended in 3 ml MACS buffer to make them pass through a 30 micrometer (µm) pre-separation filter (Miltenyi Biotech, #130-041-407) and an LS Column (Miltenyi Biotech, #130-042-401) attached to a QuadroMACS separator (Miltenyi Biotech, #130-090-976). LS Columns were washed three times using 3 ml of MACS buffer. Then, 5 ml of MACS buffer was used for flushing the CD19-positive lymphocytes with the help of the syringe provided by the LS Columns.

6.5. MinION library preparation

PCR products and amplicons were purified using ProNex beads (Promega, #NG2001) in a 1:1 ratio following the manufacturer´s protocol. The amplicons were eluted in 16 µl.

Firstly, the quality of the purified switch joint amplicons was controlled. Thus, 5 µl of purified amplicons was checked in a 0.8% Agarose (Biozym, #840004) gel. In addition, 1 µl of amplicons were loaded into 2100 Bioanalyzer (Agilent). Amplicons visible in the gel were diluted 50% before loading them into 2100 Bioanalyzer.

The outcomes of both assays were used to decide how much volume was added to each sample. First, the intensity analysis of smear bands in the Agarose gel was done (GelAnalyzer 19.1) and the values were used to make a first decision on the volume that was loaded into MinION. The intensities belonging to the same gel were considered for the following calculations:

$maxI$ = maximum intensity of the gel

$minI$ = minimum intensity of the gel

$meanI$ = mean intensity of the gel

$\Delta = maxI - minI$

Volume MinION (µl)	from	to
1	$maxI - \Delta/3$	$maxI$
1.5	$minI + \Delta/3$	$maxI - \Delta/3$
2	$meanI$	$minI + \Delta/3$
2.5	$minI + \Delta/6$	$meanI$
3	0	$minI + \Delta/6$

Volumes were then corrected by looking at the 2100 Bioanalyzer results. The correction was based on comparing samples with the same volume MinION score and seeing if they had the same pattern in Bioanalyzer. Based on the result, more or less volume was added.

The total volume of DNA amplicons must be 48 µl. The MinION library preparation (MinION Nanopore, #SQL-LSK109) followed the manufacturer´s protocol. The

only difference we made was not to use DNA control when preparing our MinION library.

6.6. Switch-joint PCR

Name	Coordinates	Sequence	Chromosome	Type	Orientation	Assembly
Sμ	105860975 - 105861001	CACCCTTGAAAGTAGCCCATGCCTTCC	14	Primer	forward	GRCh38
	106327185 - 106327211					GRCh37
Sμ	105856877 - 105856902	GGAACGCAGTGTAGACTCAGCTGAGG	14	Primer	reverse	GRCh38
	106322982 - 106323007					GRCh37
hEμ	105862197 - 105862225	GGTCACCGCGAGAGTCTATTTTAGGAAGC	14	Primer	forward	GRCh38
	106328407 - 106328435					GRCh37
Pre-hEμ	105862942 - 105862964	GGAGCCACATTTGGACGAGATGC	14	Primer	forward	GRCh38
	106329152 - 106329174					GRCh37
Sα	105588857 - 105588879	CTCAGTCCAACACCCACCACTCC	14	Primer	reverse	GRCh38
	105709106 - 105709128					
	106055194 - 106055216					GRCh37
	106175443 - 106175465					
Sγ	105626824 - 105626852	CTGCCTCCCAGTGTCCTGCATTACTTCTG	14	Primer	reverse	GRCh38
	105645549 - 105645577					
	105743825 - 105743853					
	105772166 - 105772194					
	106093161 - 106093189					GRCh37
	106111886 - 106111914					
	106210162 - 106210190					
	106238503 -106238531					
Sε	105601669 - 105601692	GGAGGGAATGTTTTT	14	Primer	reverse	GRCh38

	106068006 - 106068029	GCAGCAGCG				GRCh37
Sμ	113426178 - 113426200	GGAGGGACCCAGGCTAAGAAGGC	12	Primer	forward	GRCm38
	113389798- 113389820					GRCm39
Sα	113260804 - 113260830	GCAAGCAGTGGACCCAAAGACGAGAGG	12	Primer	reverse	GRCm38
	113224424 - 113224450					GRCm39
Sγ1	113330330 - 113330354	GCTCAGAGTGTAGAGGTCAGACTGC	12	Primer	reverse	GRCm38
	113293950 - 113293974					GRCm39
Sγ3	113361218 - 113361244	GGGCTGTTGTTGTAGCTGCAAGATAGG	12	Primer	reverse	GRCm38
	113324838 - 113324864					GRCm39
Sγ2	113307913 - 113307936	CTGATGGGGGTGTTGTTTTGGCTG	12	Primer	reverse	GRCm38
	113288916 - 113288936					
	113271533 - 113271556					GRCm39
	113252536 - 113252556					
Sε	113273349 - 113273369	GGGACCCCATTCCAGTCTAGG	12	Primer	reverse	GRCm38
	113236969 - 113236989					GRCm39

Table 6. Primers used for switch region PCR in human (GRCh38 / GRCh37) and mouse (GRCm38 / GRCm39) antibody locus.

GRCh38=Genome Reference Consortium Human Build 38, GRCh37=Genome Reference Consortium Human Build 37, GRCm38=Genome Reference Consortium Mouse Build 38, GRCm39=Genome Reference Consortium Mouse Build 39.

6.6.1. Human switch-joint PCR

Human switch polymerase chain reaction (PCR) started by mixing the different reagents as suggested for the manufacturer protocol of LongAmpTaq® (NEB, #M0323). The reaction was done in 25 µl and the optimal template amount was the isolated gDNA of 25 000 cells or less. In one reaction, the forward primer Sμ, PreSμ or hEμ was added in combination with Sα and Sγ reverse primers (Table *6*). The PCR protocol consisted of the following steps: 95°C for 3 min and 25 cycles of 95°C for 40 seconds (s), 60°C for 30s, 65°C for 3 min; and the last elongation step at 65°C for 10 min at a ramp of 4°C/s.

6.6.2. Mouse switch-joint PCR

Mouse switch PCR started by mixing the different reagents as suggested by the manufacturer protocol of LongAmpTaq® (NEB, #M0323). The reaction was done in 25 μl and the optimal template amount is 100-150 ng of isolated gDNA. In one reaction, the forward primer Sμ (*38*) was added in combination with Sα and Sγ2bc reverse primers and in another combination, Sμ in combination with Sγ3 and Sε reverse primers (Table *6*). The PCR protocol consisted of the following steps: 95°C for 3 min and 30 cycles of 95°C for 40s, 60°C for 30s, 65°C for 3 min; and the last elongation step at 65°C for 10 min at a ramp of 2°C/s.

6.6.3. Barcoded switch-joint PCR

The PCR protocols were carried out as mentioned before. PCR products used for MinION sequencing were barcoded, meaning that the primers used for the PCR contained a 20 nt barcode. The barcoded primers contain the barcode in the 5'end. The barcodes used can be seen in Table 7.

B01	AAGAAAGTTGTC GGTGTCTTTGTG	**B36**	ATGTCCCAGTTA GAGGAGGAAACA	**B72**	TAGCTGACTGTCT TCCATACCGAC
B04	TTCGGATTCTATC GTGTTTCCCTA	**B47**	GTGCAACTTTCC CACAGGTAGTTC	**B75**	GACCATTGTGATG AACCCTGTTGT
B05	CTTGTCCAGGG TTTGTGTAACCTT	**B49**	ACTGGTGCAGCT TTGAACATCTAG	**B76**	ATGCTTGTTACATC AACCCTGGAC
B09	AACTAGGCACA GCGAGTCTTGGTT	**B51**	GTTGAATGAGCC TACTGGGTCCTC	**B78**	AACAACCGAACCT TTGAATCAGAA
B12	CAGGTAGAAAGA AGCAGAATCGGA	**B55**	TGGAAGATGAGA CCCTGATCTACG	**B81**	CCTCATCTTGTGA AGTTGTTTCGG
B17	ACCCTCCAGGA AAGTACCTCTGAT	**B60**	CATGTTCAACCA AGGCTTCTATGG	**B92**	TTGTGAGTGGAAA GATACAGGACC
B24	GCATAGTTCTG CATGATGGGTTAG	**B64**	GACAGACACCGT TCATCGACTTTC		
B32	CCAGTAGAAGT CCGACAACGTCAT	**B68**	GAATCTAAGCAA ACACGAAGGTGG		

Table 7. The sequence of barcodes in 5'- of switch primers.

6.7. Peyer´s patches isolation

Intestines from dead mice were obtained in cold PBS-T (Biotek, EL406). Using a 1 ml syringe (BRAUN, #91614060) and a needle (Fine-Ject, #4710006030), the insides of the intestines were washed with PBS-T. A tube was introduced into the intestines for better inspection.

Peyer's patches (PP) were identified as a white bulked mass on the surface of the intestines. Isolation of PP was performed using scissors. Isolation of B cells from PP was done by smashing the cells against a filter (Miltenyi Biotech, #130-041-407), centrifuging the content and washing the cells twice using PBS before continuing to isolate gDNA.

6.8. Statistics

Tests were two-tailed, and a p-value lower than 0.05 was considered statistically significant. Shapiro-Wilk test was used for normality testing of continuous variables, using R function `shapiro.test()`. An independent Student t-test was used to compare two groups when continuous data met the criteria for normality, using the R function `t.test()`; otherwise, the Mann-Whitney U test was used, using the R function `wilcox.test()`. Comparison of multiple groups was made using ANOVA post-hoc analysis for data that met the criteria for normality, using the R `aov()`; otherwise, the Kruskal-Willis test was done, using the R `kruskal.test()`. Principal component analysis (PCA) was performed with scaled data and under the same seed using R (seed=147). PCA was performed using the R function `prcomp::prcomp()`.

6.9. SWIBRID

Firstly, SWIBRID demultiplexes the samples from MinION using the 20 nt barcodes (Table 7). Then, the reads that did not meet our quality criteria were discarded. Our quality criteria were the following:

- Reads are longer than 500 nucleotides (nt)
- Reads contain the SM forward (FW) primers and the SG reverse (RV) or SA RV within the 100 nt beginning and end of the read
- If two primers are found together in the middle of the read, the read should be split and consider them as two independent reads
- Reads should not contain single primers outside the 100 nt of the beginning and end of a read
- The sequence of the read must contain switch regions (Table 8)
- The sequence of the read must align in more than 95% of its length to the genome
- If more than 50 nt of the sequence does not align to switch regions, it will be considered a J-CH1 insert as long as its sequence aligns in 95% of its length to the genome.
- Next, the filtered reads were clustered by similarity, specifically by calculating the Cosine distance between the cleaned coverage patterns. The distance measure compared how similar all reads are to each other using the coordinates of alignment and generated a phylogenetic tree. As a result, a dendrogram of read differences was generated for the cut-off determination.

Name	Specie	GenBank or reference	Length (bp)
Switch mu	Human	X54713.1	4 525
Switch gamma 3	Human	U39935.1	3 395
Switch gamma 1	Human	U39737.1	4 171
Switch alpha 1	Human	L19121.1	3 326
Switch gamma 2	Human	U39934.1	3 812
Switch gamma 4	Human	X56796.1	4 757
Switch alpha 2	Human	AF030305.1	2 758
Switch mu	Mouse	AF446347.1	870
Switch gamma 3	Mouse	D78343.1	1 801
Switch gamma 1	Mouse	M12389.2	5 188
Switch gamma 2b	Mouse	(*142*)	4 176
Switch gamma 2c	Mouse	(*142*)	2 104
Switch alpha	Mouse	D11468.1	3 578

Table 8. Switch region consideration to align for SWIBRID

The cut-off limits the differences allowed to take part in cluster determination. In order to determine the cut-off, the function of clusters and cut-off is used. In this function, two tendencies were observed, i) one with a high slope observed at low cut-offs, and ii) one with a lower slope observed at larger cut-offs. The middle point between these two slopes was used to determine the cut-off and, thus, the number of clusters.

7. RESULTS

DNA repair malfunction predisposes people to develop diseases such as cancer and it is often diagnosed by analyzing mutations in DNA repair proteins related to malfunction. Such an option of diagnosis is chosen because researching somatic DNA repair in cells is complex due to the low frequency and randomness of somatic DNA breaks in cells. However, mutation screening of DNA repair proteins does not directly study the efficiency of DNA repair. Therefore, we decided to focus on researching DNA repair function by studying somatic breaks by analyzing genomic scars in the antibody gene during B cell diversification. Among the B cell diversification events is CSR, a controlled deletional recombination process in the antibody locus that joins two distant switch regions to give rise to a new isotype generating a switch-joint. Thus, switch-joints result from a DNA repair event and commonly carry mutations and insertions serving as an analyte to study DNA repair footprints.

On the other hand, non-immunoglobulin elements deriving from distant genomic regions (inserts) integrate into the heavy chain antibody locus (IGH), thereby adding another layer of diversity to the antibody repertoire (*93*, *94*). More specifically, inserts incorporated in the switch regions (J-CH1 inserts) can be spliced into antibody transcripts and consequently incorporated as extra elements in the antibody protein (Figure *6*A). The mechanism of J-CH1 insert acquisition is still unclear. Recent studies identified J-CH1 inserts in antibody transcripts with a mean length of 160 bp and were even estimated to comprise up to 10 kilobases (Kb) in the genome when counting their multiple exons separated by large introns. In this study, the J-CH1 inserts were studied using suppression PCR, an improved PCR to enrich the amplification of longer amplicons (*143*). Unfortunately, suppression PCR removed the possibility of quantifying inserts per clone since such amplification is biased towards the amplification of long sequences (*97*). An unbiased approach to identify J-CH1 inserts would allow the quantification of B cell clones and, therefore, the J-CH1 insert frequency (J-CH1 insert per B cell clone). Also, the study identified J-CH1 inserts spliced in the transcripts, leaving aside those insertion events that were spliced out but still happened and are essential to account when studying the J-CH1 insert mechanism. In other words, the J-CH1 insert mechanism should be studied at the DNA level using a tool to identify the insertions and unique B cell clones for a quantitative analysis.

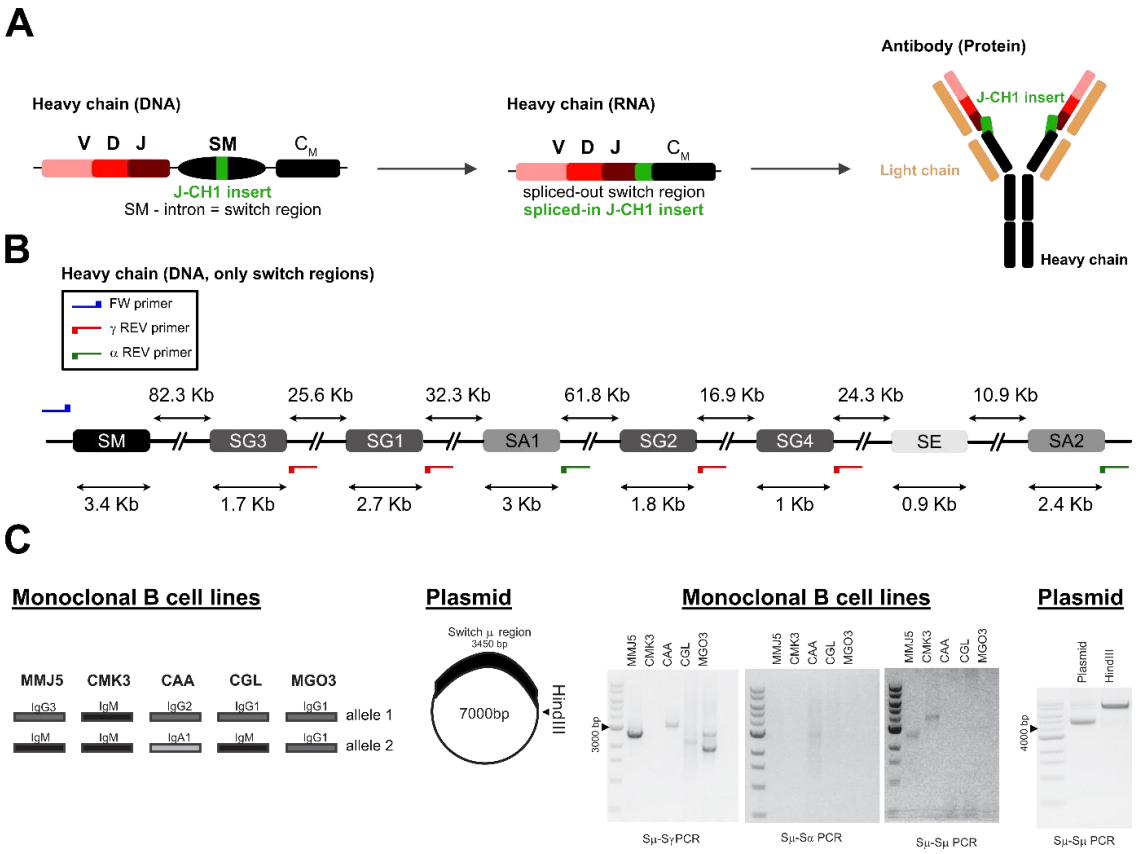
The switch-joints have been previously studied using short sequencing (Illumina or Ion Torrent), which captured only a few nucleotides covering the switch-joints (*144*). However, our idea aims to study a bigger picture of the switch-joints using long-read sequencing (L-NGS) MinION to screen CSR breakpoints to identify sequential switching, intra-switch deletions and J-CH1 inserts in long switch-joints reads, among other aspects.

Thus, this thesis aims to develop a tool to analyze switch-joints and characterize their variabilities (J-CH1 inserts and breakpoints) within a library to determine unique B cell clones using long amplicons derived from MinION.

7.1. Development of a bioinformatical tool to identify breakpoint profiles, unique B cell lines and J-CH1 inserts using long reads

7.1.1. Development of a tool that overcomes MinION errors in the sequence and identifies individual B cell clones in a sample

Building upon previous work (*94*), I first aimed at improving the methodology that integrates i) amplification of switch-joints comprising the switch μ and switch γ 1, γ 2, γ 3, γ 4, α 1 or α 2 regions (SM, SG1, SG2, SG3, SG4, SA1, SA2) (*94*), ii) sequencing by MinION technology and iii) computational analysis capable of quantifying B cell clones via switch-joint analysis and J-CH1 insert. The primers were designed to amplify only switch-joints, e.g., SM-SG3, while IGH that did not undergo CSR were excluded from amplification since the smallest distance found between our FW primer and any RV primer in the IGH locus that was not subjected to CSR is 87.4 Kb (SM-SG3), which is far beyond any PCR amplification limit (Figure *6*B) (*145*).



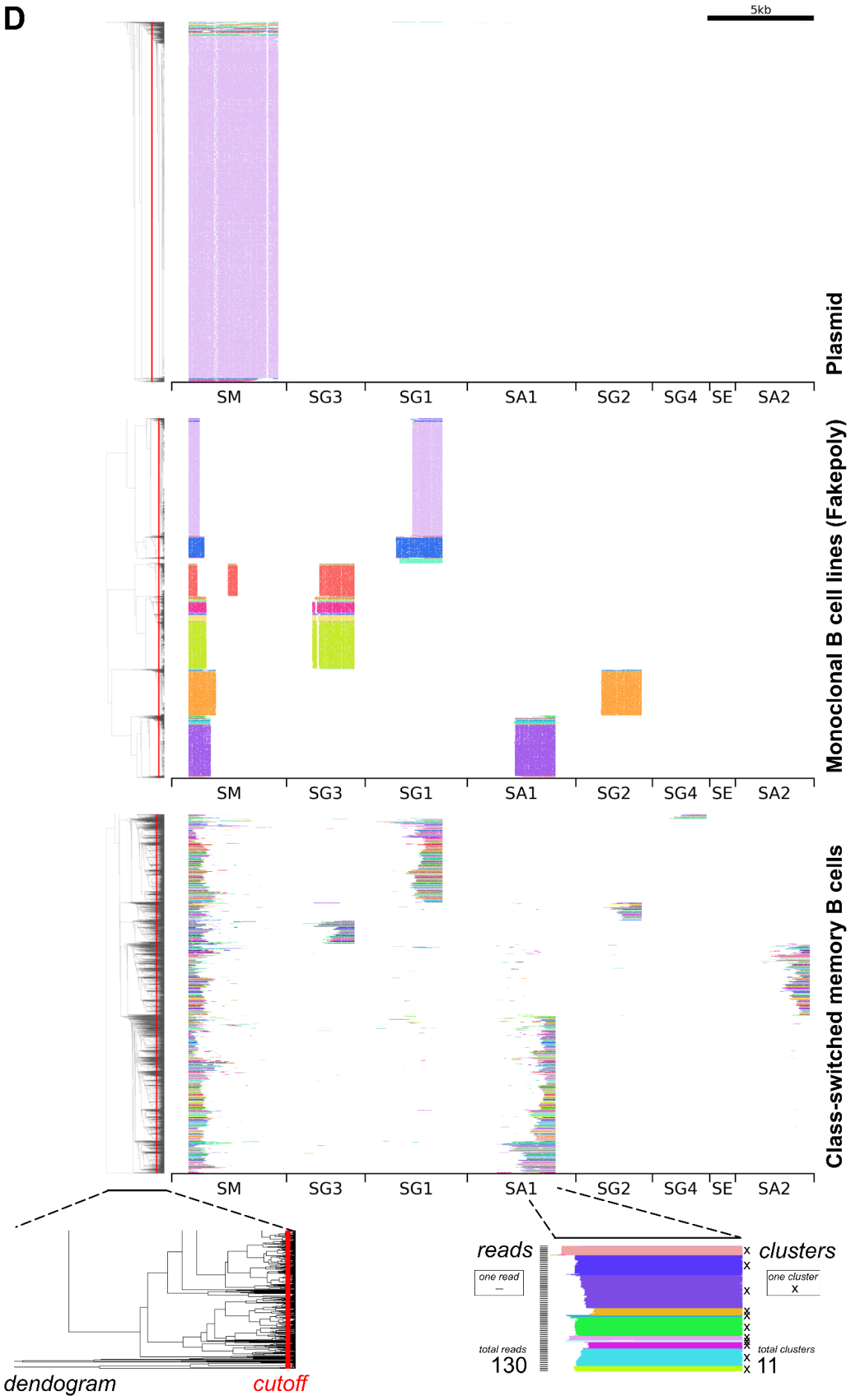

D
5kb
Plasmid
Monoclonal B cell lines (Fakepoly)
Class-switched memory B cells
SM
SG3
SG1
SA1
SG2
SG4
SE
SA2
dendogram
cutoff
reads
one read
total reads
130
clusters
one cluster
total clusters
11

Figure 6. SWIBRID: strategy and outcome of the switch-joint analysis.

A. Depiction of the immunoglobulin heavy chain (IGH) gene, transcript and protein of a J-CH1 insert-containing antibody. C_M=constant μ, SM=switch μ region. J-CH1 insert=green, V, D, J domains=red and SM & C_M=black. **B.** Depiction of the switch regions of the human IGH locus indicating primer binding sites to amplify switch-joints. Kb=kilobases. SG3=switch γ3 region; SG1=switch γ1 region; SA1=switch α1 region; SG2=switch γ2 region; SG4=switch γ4 region and SE=switch ε region and SA2=switch α2 region. Primers: SM forward (FW)=blue, SG reverse (RV)=red, SA RV=green. **C.** On the left side, the so-called "FakePoly" control comprises 6 switched alleles of monoclonal B cells, deriving from EBV immortalized human B cells, namely, MMJ5, CAA, CGL and MGO3 (CMK3 acts as a negative control because it does not carry a switched allele); a plasmid containing the human SM used as control after linearization using restriction enzyme HindIII. On the right side, 0.8% agarose gels show the switch-PCR of the FakePoly or SM-SM PCR of the plasmid. A 1 Kb ladder was used. **D.** Read plots of samples with no-, medium- and high-B cell diversity, namely, plasmid < Fakepoly < class-switch memory B cells. Five thousand random reads are depicted in each plot. On the lower left, a hierarchical clustering dendrogram of the switch-joint reads is shown together with the cut-off (red) used to determine the number of clusters that discriminates technical noise from biological diversity. On the lower right, each straight line in the same horizontal plane represents an individual read or switch-joint amplicon. Reads depicted together and in the same color belong to the same B cell cluster. The x-axis represents the IGH locus switch regions and the reads only appear when aligning with them.

Switch joint amplification (switch-PCR) generates long amplicons with a mean length of 2.5 Kb. Long amplicons can be sequenced using L-NGS devices, e.g., MinION or PacBio, but also by short amplification, e.g., Illumina or Ion Torrent, upon fragmentation of the amplicons. The latter generates highly accurate reads with a maximum read length of 2x300 bp. However, short reads demand reconstitution and assembly of amplicons using computational tools. Considering that switch regions are characterized by long, repetitive GC-rich sequences (*142*), accurate reconstitution of fragmented switch-joint amplicons is excluded. Thus, we decided to sequence our amplicons using the L-NGS device, and there were two options to consider: MinION from Nanopore technology® and PacBio from Pacific Biosciences®. PacBio sequencing gives rise to more accurate reads than MinION (error rate <1% versus 6.6%, respectively), but it is also known to provide a lesser read yield (PacBio 8 gigabytes in contrast to 50 gigabytes of data per flowcell of MinION). MinION is a more flexible technology than PacBio, due to the possibility of sequencing without the need for PCR amplification (*146–148*). It is also a technology that has rapidly evolved to acquire high read yields (in 2015, 16 000 reads and in 2021, 48 million reads), fewer errors (in 2015, 12% of errors and in 2021, 6.6% of errors) and new features (specific cytosine DNA methylation identification) (*149–152*). Betting on the technology that has shown constant improvements in recent years, we decided to use MinION despite its

higher error rate, which we tackled using a bioinformatical approach (see section 7.1.3).

In a previous study, J-CH1 insert frequency was standardized using total reads, which in turn meant that every read was considered one B cell clone (*94*). Such an approach increases the risk of making wrong assumptions about J-CH1 insert frequency because unique B cell clones are represented by several reads in polyclonal B cell libraries. Therefore, my thesis aims at developing a tool to identify individual B cell clones for a proper J-CH1 inserts frequency quantification. To this end, we established a collaboration with the bioinformatician Dr. Benedikt Obermayer (Berlin Institute of Health) to translate my biological idea of how the tool should identify J-CH1 inserts and individual B cell clones and characterize the breakpoints in switch-joints into an algorithm using his deep bioinformatical skills. Together, we developed a bioinformatical tool called SWIBRID (SWItch region Beakpoint Repertoire IDentification), the first computational tool to identify unique B cell clones using switch-joints. SWIBRID:

- Demultiplexes libraries in samples with the help of barcodes (Table 7),
- Discards reads that do not meet our quality criteria (section 7.1.2),
- Identifies:
    - Breakpoints of switch-joints,
    - J-CH1 inserts,
    - Isotypes
    - Characterize switch-joint scar-features,
- Classifies the switch-joint amplicons in individual B cell clones (section 7.1.3),
- Characterizes the switch-joint B cell clones by determining the number of reads that represent them.

Establishing controls to ensure SWIBRID performance in B cell clustering by switch-joint analysis was crucial. SWIBRID read input contains errors due to PCR amplification and MinION sequencing, which can add technical noise to the reads library and potentially lead to overestimating clones. Therefore, two samples were generated and established as a control to check for i) PCR amplification by amplifying switch-joints of well-defined B cell monoclonals pooled together that we called "FakePoly" (Figure *6*C); and ii) MinION by sequencing a linearized plasmid containing the SM region using MinION (Figure *6*C).

To visualize the output of SWIBRID comprehensively, we developed the readplot. The readplot visually shows the clustering of a library for a fraction of randomly chosen reads (Figure *6*D). The readplot represents the read's alignment to the IGH locus, using the x-axis to reference the switch regions (Table 8). For this purpose, the x-axis depicts the scaled SM, SGs, SE and SAs regions in the same order found in the IGH locus. Exons and intergenic regions of the IGH locus were excluded from the x-axis. The y-axis indicates the individual reads. Every read starts in SM since we use the SM FW primer at the very left side of the plot and ends in an SG or SA region. Consequently, sharp edges can be found at the beginning of SM and the end of every other switch region. Reads are represented by fine, straight lines depicted from left to right and are only visible over the switch

region parts to what they align. Reads considered to belong to the same B cell clone are depicted together and in the same color; thus, the diversity of a B cell sample is represented by the variety of colors in a readplot.

The readplots serve as a clustering visualization tool to judge the sample's diversity. Visual comparison of readplots comprising 5 000 reads of plasmid (low diversity), FakePoly (medium diversity) and class-switched memory B cells (high diversity) show clear differences. Plasmid portrays a unique big cluster, FakePoly, a few big clusters and the class-switched memory B cells several small clusters (Figure *6*D). More specifically: i) in plasmid, 95% of the reads fall into one cluster, in line with the fact that it is supposed to contain a single B cell clone carrying SM; by contrast, ii) FakePoly contains 6 clones that comprise 84% of all reads, and the clones have an average of 704.5 reads, 209-fold more reads than the rest of clusters; iii) the class-switch memory B cell sample has 2 255 clones; from them, 29% are singletons (clusters with one read). Interestingly, the biggest cluster is formed by only 28 reads (unlike for plasmid and Fakepoly, 4 758 and 1 578 reads, respectively), which shows how well-balanced the reads among clusters are in the class-switched memory B cell sample (Figure *6*D). The high amount of singletons is a SWIBRID's limitation, which we overcame using what we call the effective number of clusters (eff_nclusters) (section 7.1.3).

In summary, we identified individual clusters from a switch-joint read library produced by MinION using our computational tool SWIBRID. Such an analysis gave rise to the readplots, a clustering visualization tool that allows for the sample's diversity judgment.

7.1.2. The development of SWIBRID: the tool that characterizes switch-joints

The success in clustering polyclonal B cell samples using switch-joints was a joint effort culminating in the development of SWIBRID. SWIBRID's structure consists in i) filtering the reads that do not match our quality criteria, ii) identifying J-CH1 inserts, iii) clustering the reads by measuring the differences between reads and iv) determining the number of individual B cell clones in a sample (Figure 7A).

Raw reads were first filtered to obtain a high-quality library to be clustered. Our quality criteria involved discarding reads with the following properties: i) reads shorter than 500 bp (short), considering that we were aiming for an extended range sequence analysis; ii) reads not containing primers at both ends of the reads (incomplete), hypothesizing that the sequencing of the read has been abrogated during the MinION run, iii) reads that contain primers in the middle (internal), iv) the reads in which less than 95% of the sequence map to the genome (low_cov) and v) reads that do not contain any switch region (Table 8) (no-switch), considering that we were interested in switch-joints (Figure 7A). Filtering MinION reads was crucial because PCR amplification could create artifacts and generate false diversity. Indeed several artifacts were observed in a manually performed analysis. An early version of SWIBRID identified 37.6% of reads as internal, discarding almost half of the libraries. These reads contained

a primer in the middle of the read, and we decided to discard them (Figure 7B-top). Manual analysis of internal reads revealed that they mostly consisted of two individual reads joined by their primer ends. The high amount of internal reads could have resulted from the adaptor ligation step during the library preparation (SQK-LSK109) for MinION, which uses T4 ligase to ligate an adapter to every amplicon for the sequencing to occur. We decided to split internal reads through the internal primers and improved our analysis by increasing the number of mapped reads (>95% of the sequence of a read maps to the genome), decreasing the number of internal reads, and in the aftermath, gaining eff_nclusters (Figure 7B-below) (section 7.1.3). During the filtering process, we needed to identify J-CH1 inserts by differentiating between no-switch reads and reads containing part of switch regions plus other parts of the genome since the latter could be a J-CH1 insert-containing read. Following previous studies in identifying J-CH1 inserts in switch-joints (*94*), we decided to identify J-CH1 inserts as a sequence belonging to a read of at least 50 nt long flanked by switch regions and which sequence aligns at least 95% to the genome. Manual analysis of the J-CH1 insert-containing read confirmed the accuracy of SWIBRID identification (Figure 7D). In summary, the filtering step discarded short, incomplete, low_cov, internal and no-switch reads and identified J-CH1 inserts in parallel.

Next, the filtered reads were clustered by similarity, specifically by calculating the distance measure. The distance measure compares how similar all reads are to each other using the coordinates of alignment and generates a phylogeny tree. The distance measure in SWIBRID combines two measurements: i) the interval distance that considers the intersection and union of alignment blocks (Cosine distance) and ii) the switch overlap distance, which compares coverages of each of the switch regions in reads. Both are calculated and weighted equally to calculate the differences between the reads in a sample. As a result, a dendrogram of read differences is generated (Figure *6*D-left below).

The last step is to identify the number of individual B cell clones. In order to do so, we need to determine the cut-off. The cut-off sets the limit of differences and allows identifying unique clusters (considered B cell clones). The cut-off aims at separating technical noise from biological diversity. To visualize the diversity of samples, we plotted the number of clusters (y-axis) obtained at varying cut-offs (x-axis). Of note, two clear tendencies were observed. For very small cut-offs, the graph showed a steep slope representing the technical noise. Beyond a certain cut-off, another tendency could be shown with a lower slope representing the biological diversity (red and blue in Figure 7A-low right, respectively). As a result, the cut-off was determined by the intersection between the technical noise and the biological diversity. The cut-off can be seen represented in Figure *6*D-left below as a red line crossing the dendrogram. When counting the lines that the cut-off intersects, we obtain the number of clusters that the samples have. The clusters obtained at the calculated cut-off were called nclusters.

SWIBRID aims to represent B cell diversity through switch-joints by only considering events that would appear in nature. In order to achieve this, SWIBRID follows the following B cell class-switch biology rules:

- A switch-joint can include more than two different switch regions. B cells can class-switch more than one time due to a process called sequential switching (*73*)
- Class-switching can only occur following the (-) sense direction of the IGH locus due to the intragenetical deletional recombination. In other words, if the SM-SG1 joint occurs, SG3 should not appear after SG1 (Figure *6*B), or it would be considered a J-CH1 insert (*153*)

SWIBRID has been constantly optimized through manual analysis of SWIBRID output and weekly discussions. One significant problem we experienced was the case of an extensive set of reads that did not contain barcodes (undetermined reads). Those reads occupied more than 35% of the library in thirteen independent MinION runs, when usually a sample occupies only up to 10% of a run (Figure 7C). Twenty reads from the undetermined reads were manually analyzed and identified as the control DNA used by the library preparation SQK-LSK109 Kit from Nanopore technology®. Since then, the control DNA was excluded from the library preparation and the libraries were more enriched in the sample reads.

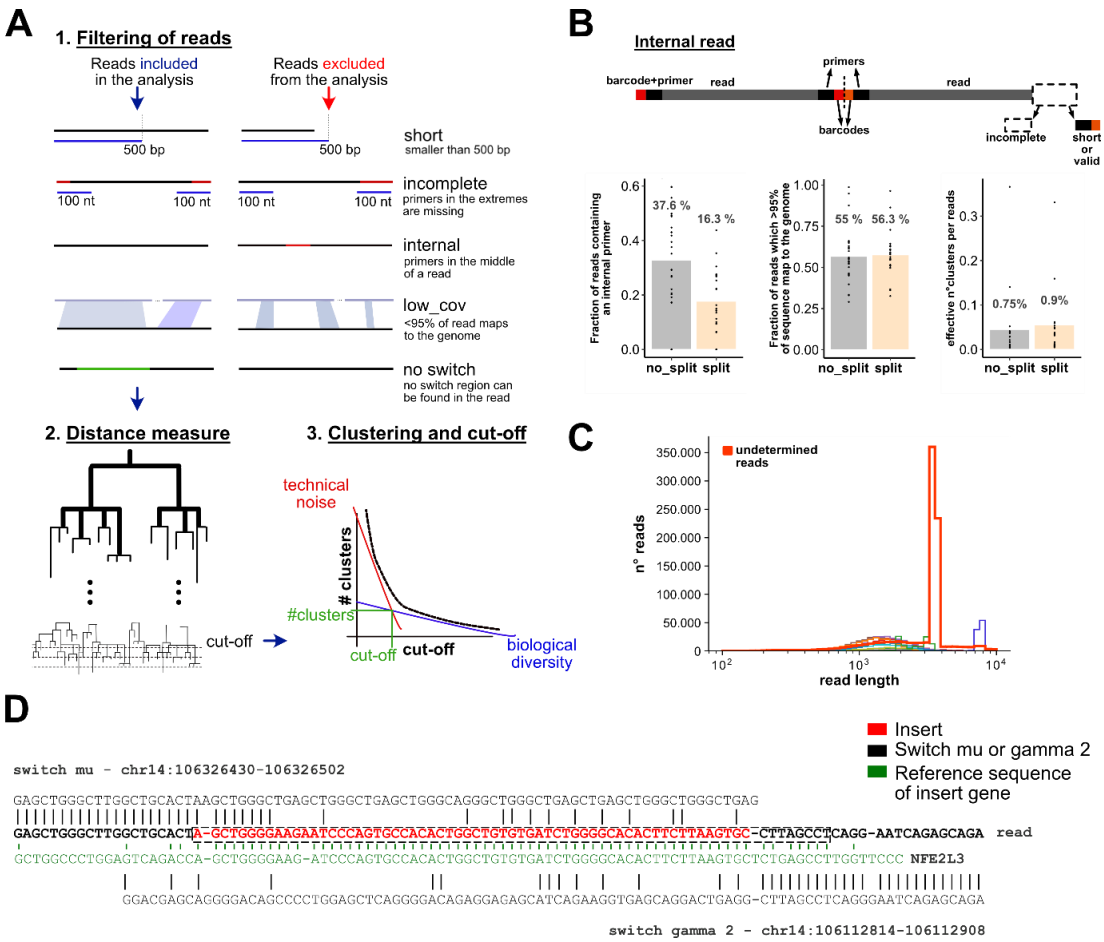


Figure 7. Towards the optimization of SWIBRID.
A. Simplified scheme of SWIBRID's pipeline. First, a filtering process to exclude reads not fulfilling our quality criteria is performed (1), followed by the measurement of read distances (2) and last, the determination of cut-off, which takes part by differentiating the technical noise from the biological diversity to achieve the number of clusters that are representative for the sample number

of B cell clones (3). B. Manual analysis of internal reads. At the top is a scheme of the overview of an internal read. At the bottom, plots representing the consequence of splitting reads in, from left to right, internal reads, reads mapping to the genome and the number of effective clusters, no_split_n=21, $split_n$=21. The value in percentage indicates the median. C. Graphical representation of a library of reads after demultiplexing (classification by barcodes). In orange is the set of undetermined reads. The rest of the colors represent reads that contain barcodes. D. A representative manual analysis of a J-CH1 insert-containing read identified by SWIBRID. The bold part of the read contains the J-CH1 insert. The part of the sequence identified as a J-CH1 insert by SWIBRID is in red, while the sequence highlighted by a dashed box was identified as a J-CH1 insert by the manual analysis. NFE2L3 insert reference sequence, in green, corresponds to chr7:26198404-26198500. Coordinates belong to GRCh37 genome assembly.

In summary, SWIBRID's structure follows highly optimized biological and computational rules, making it possible to identify J-CH1 inserts and individual B cell clones using switch-joints generated by MinION.

7.1.3. Analysis of naturficial reads and human polyclonal B cell samples show that the read numbers do not impact SWIBRID output in highly diverse samples

In the comparison between read numbers and the number of B cell clones identified by SWIBRID in samples, we hypothesized that the number of reads might have an impact on the nclusters identified in a sample: fewer reads, fewer clones. If that were the case, comparing results from samples with a different number of reads could bias the results. To address this problem, we decided on two things: i) generate artificial switch-joints representing different amounts of clones represented by various numbers of reads and ii) analyze human samples at different cell numbers.

Firstly, I generated a script to produce artificial switch-joints that I named the "*in-silico* prototype." It generated artificial reads to represent switch-joint libraries based on the switch region annotations. Individual B cell clone-like reads were randomly created following four pipelines: Deletion, Random, Consecutive 1 and Consecutive 2. i) Deletion generated joins between SM and Sx between two random points. In the join, SM carried a deletion of a length between 100 and 700 nt in a random position before switching. ii) Random is a pipeline similar to Deletion because it generates a join between SM and Sx at random positions, but in addition, it creates a deletion of random size in both SM and Sx. iii) Consecutive 1 generates a join between SM, SX and SXX at random positions mimicking a common event in CSR called sequential switching (*73*). iv) Consecutive 2 follows the same principle as Consecutive 1, but SM switches thrice to different switch regions using random positions (SX, SXX and SXXX) (Figure 8A).

The *in-silico* prototype successfully generated the desired amount of completely distinct reads. However, the artificial reads generated did not represent the usual distribution of breaks or isotypes in MinION libraries of a natural B cell repertoire. The mean length of artificial reads was 4 388 kb, while the mean length of real samples was 1 367 kb. The cause was that the *in-silico* prototype did not account for the frequency of breaks occurring in natural CSR events. We addressed this issue by expanding the *in-silico* prototype to the generation of naturally templated synthetic reads, which we called naturficial reads.

The templates of the naturficial reads were based on the coordinates of previously identified B cell clones by SWIBRID, ensuring the natural distribution of isotype switching and breakpoint acquisition in the newly generated reads. In detail, B cell clones were produced using the reference sequence of nclusters found in eight samples from four donors and later mutated with the corresponding degree and quality of mutations that were recorded in the SM plasmid control, which represented the mutations or errors that MinION produced (Figure *8*B) (Table 5). The minimum and the maximum number of individual naturficial B cell clones generated were 10 and 10 000, respectively. Naturficial reads are considered an *in silico* library.

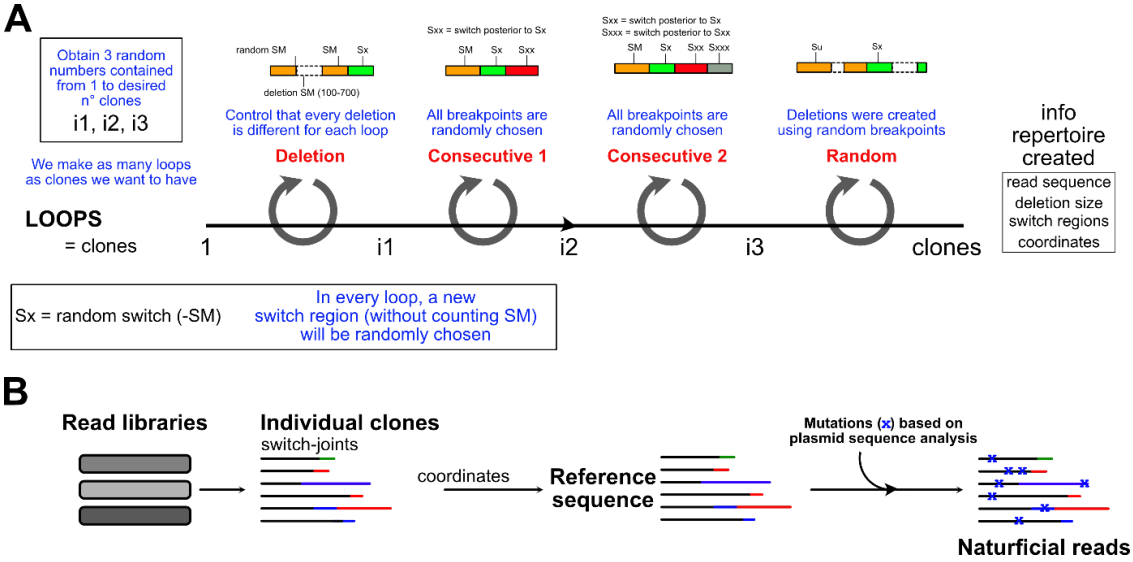


Figure 8. Artificial and naturficial reads.

A. The scheme depicts a loop-based script to generate randomized artificial reads representing unique switch-joints of B cell clones. Four types of artificial reads were generated: Deletion, Consecutive 1, Consecutive 2 and Random. The uniqueness of each clone was confirmed after every loop. **B.** Scheme of naturficial read generation by Dr. Benedikt Obermayer. The black line represents SM and the other colors SAs or SGs. Blue “X” represents a mutation from Table 5.

Secondly, to analyze human samples, three healthy donors (HD) were analyzed in triplicates using different cell numbers to test SWIBRID reproducibility

considering potential technical variabilities (e.g., read number, PCR amplification, MinION flowcell sequencing). This experiment will be referred to as the cell number experiment. Primary human B cells were isolated, and memory IgG and IgA ($CD27^+IgG^+/IgA^+$) were sorted following the gating strategy in Figure *9*A ($CD27^+IgG^+/IgA^+$ in total B cells from HD1: 10.8%, HD2: 14.5% and HD3: 6.5%). Based on the quality between the input and output in a PCR reaction, we identified 500, 5 000, 10 000, 25 000, 50 000, 100 000, 150 000 and 200 000 memory IgG and IgA B cells as the most suitable samples to be tested. I expected to obtain a linear relationship between the input cells and the output B cell clones from SWIBRID. PCR reactions using these cell numbers were performed in three technical replicates for each donor. The isolation of genomic DNA (gDNA) of the three HD and the switch-joint PCR was performed following the same protocol. Before sequencing, the switch-joint PCR result was checked by agarose gel and Bioanalyzer (Figure *9*B-C).

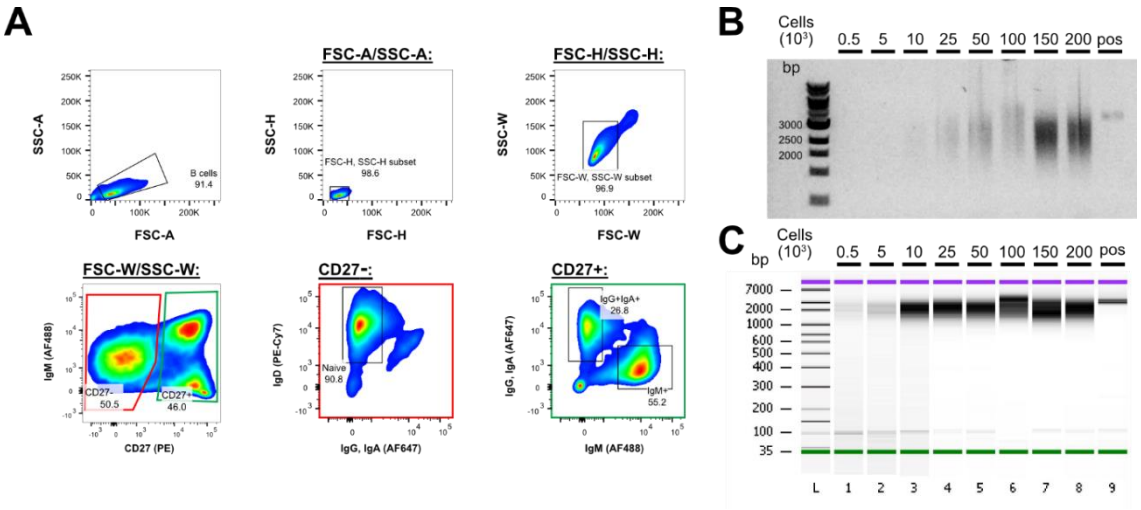


Figure 9. Preparation and quality control of samples for the cell number experiment.

A. Representative FACS gating strategy for sorting. Sorting was performed using FACSAria Fusion Flow Cytometer (BD) and was analyzed by FlowJo software. The starting material was primary human CD19+ B cells pre-sorted via magnetic bead separation. CD27-PE, IgM-AF488, IgG-AF647, IgA-AF647, IgD-PE-Cy7. AF=AlexaFluor. **B.** 0.8% agarose gel showing 30% of amplicons generated by the switch-joint PCR. **C.** DNA analysis by chip-based capillary electrophoresis (Bioanalyzer, Agilent) of 3-6% of the amplicons generated by the switch-joint PCR. Pos=positive control from EBV-immortalized IgA expressing monoclonal B cell. In green and magenta, the limit markers of Bioanalyzer.

SWIBRID analysis of the cell number experiment and naturficial reads comprised identifying unique switch-joint B cell clones using two units: nclusters and eff_nclusters. As previously described, nclusters are defined by the number of unique switch-joint B cell clones found in a sample at a defined cut-off (Figure

7A). Considering that our analysis is based on B cell biology and that MinION has a 5% error rate, we decided to ignore small nclusters and look at the data in a more realistic way giving rise to eff_nclusters. Thus, eff_nclusters is defined by the smallest number of nclusters containing 95% of the reads (Figure *10*A). We chose this approach following the VDJ-sequencing B cell repertoire analysis, which discards reads or clusters under specific quality rules (*154*). It is essential to differentiate these two units since nclusters contain singletons (clusters that only have one read), and eff_nclusters contain a more considerable percentage of clusters represented by more than one read.

SWIBRID analysis of the cell number experiment and naturficial read libraries identified an amount of B cell clones that increased with the input number of cells or clones, respectively. There was almost no difference between nclusters and eff_nclusters in the three HD of the cell number experiment, as shown by the correlation index, R2. In both cases, HD showed a linear tendency between input cells and SWIBRID switch-joint B cell clones (Figure *10*B). However, the maximum amount of nclusters and eff_nclusters found were 27 000 and 26 000, respectively, even though the input amount of cells for these values was 100 000. The naturficial reads only used a maximum amount of input clones of 10 000, which in principle, and based on the cell number experiment, SWIBRID should identify efficiently. The eff_nclusters identified in naturficial reads followed an intersected line (dashed line in Figure *10*C) that showed an equal number of SWIBRID and input clones, which means that SWIBRID could successfully identify the exact amount of input clusters (Figure *10*C-right). On the contrary, the tendency of nclusters diverged from the intersected line at low and high numbers of input clones (Figure *10*C-left). The difference in the values obtained of nclusters and eff_nclusters in naturficial reads could be due to the impact of the MinION mutations. MinION mutations could falsely create unique switch-joint B cell clones that would not cluster together with other reads, creating singletons. Therefore, the usage of eff_nclusters is preferred over nclusters. Thus, SWIBRID can identify B cell clones successfully using eff_nclusters and the maximum eff_nclusters identified should not be higher than 20 000. Since the number of cells does not always reflect the number of clones, analyzing a sample represented by various amounts of cells would benefit the analysis of switch-joints.

We next tested the hypothesis that the number of reads can affect the output number of SWIBRID clones. We analyzed the different input clones of the naturficial reads using 10 000, 25 000, 50 000, 100 000 and 200 000 reads. nclusters, but not eff_nclusters, identified in naturficial reads increased with the number of reads used in the sample (Figure *10*C). The cell number experiment gave rise to different amounts of reads; however, we did not observe evidence of an affected analysis. The results supported the usage of eff_nclusters over nclusters since it was not subjected to the number of reads in the naturficial reads.

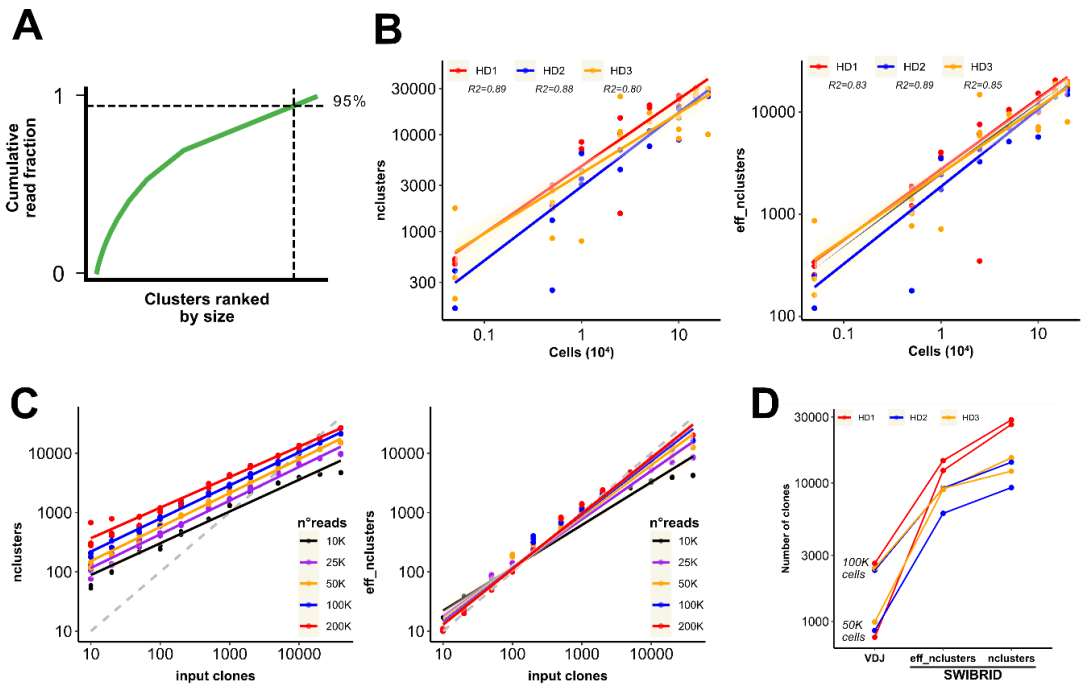


Figure 10. Effective clusters identify individual B cell clones after SWIBRID analysis of *in vivo* and *in silico* samples.

A. Explanatory scheme of identifying eff_nclusters, which corresponds to 95% of the biggest clusters in a sample. X-axis clusters are ordered by size, and the y-axis is the cumulative number of reads. **B.** Human healthy donor (HD) samples run through SWIBRID. The input of IgG^+/IgA^+ B cells is represented by the x-axis, while the output number, given by SWIBRID, of nclusters (left) and eff_nclusters (right) are represented on the y-axis. Three technical replicates were done for each cell number sample and donor. The lines represent the mean of the replicates per donor. HD1 and HD2 have three replicates per cell amount, HD3 has three replicates in 500, 5 000, 25 000, 150 000, 200 000, 100 000 and two replicates in 10 000, 50 000. **C.** Naturficial reads run through SWIBRID. The input clones are represented by the x-axis, while the output number, given SWIBRID, of nclusters (left) and eff_nclusters (right) are represented on the y-axis. Colors represent the different number of reads comprising each library: 10K=black, 25K=purple, 50K=yellow, 100K=blue and 200K=red. K=1 000. The dashed grey line is the function x=y. **D.** Comparison of B cell clones obtained by VDJ-sequencing and SWIBRID using 50 000 and 100 000 memory $IgGA^+$ B cells of three donors. Points show the mean of three technical replicates. The three donors are represented by different colors: HD1=red, HD2=blue and HD3=yellow.

SWIBRID results showed that the three HD did not elicit more than 27 000 unique switch-joint B cell clones even though the input cell triplicated that value (Figure *10*B). Since our method was novel, we decided to gain further insights to understand if SWIBRID was underestimating the clonality of our samples.

Currently, the only method to analyze B cell clonality is VDJ-sequencing. VDJ-sequencing identifies unique B cell clones based on the unique combination of V-D-J-constant domains (VDJ-clone) (*155*). Considering that the saturation of SWIBRID B cell clone identification begins at 50 000 cells, we ran VDJ-sequencing in the three HD using 50 000 and 100 000 cells as inputs. VDJ-sequencing revealed that the three HD had a maximum of 1 000 and 3 000 VDJ-clones for 50 000 and 100 000 input cells, respectively (Figure *10*D). Therefore, SWIBRID might not be underestimating the B cell clones identified. The vast discrepancy between the unique B cell clones found in the VDJ-sequencing and SWIBRID can be explained by the fact that they identify different types of clones. While VDJ-sequencing differentiates B cell clones generating antibodies binding to different antigens, SWIBRID differentiates between B cell clones that have undergone different amounts of CSR rounds. Thus, one VDJ-clone can be covered by several SWIBRID clones (eff_nclusters).

In summary, naturficial reads showed that by identifying the eff_nclusters, we avoided the additional diversification caused by MinION mutations and the different amount of reads. We also can conclude that VDJ-clones are not equal to SWIBRID clones since they represent antigen specificity and B cell diversification, respectively.

7.2. Translation of the human pipeline to murine B cells: *in vivo* and *in vitro* models of mouse switch-joint diversity

We successfully generated a workflow to analyze human samples using SWIBRID. The development of SWIBRID aimed to analyze the J-CH1 insert incorporation by modulating DNA repair proteins *in vitro*. Nevertheless, the *in vivo* analysis of samples carrying DNA repair protein deficiencies is also crucial for understanding the impact of DNA repair proteins in the natural J-CH1 acquisition. Human DNA repair protein deficiencies are rare diseases (*156*, *157*), implying that procuring such material is limited. In fact, most of the research is done on DNA repair protein deficiencies in mice than in humans. Therefore, we found that translating our pipeline to the murine system would only wide-up the research possibilities.

To do so, we collaborated with Dr. Michela Di Virgilio's laboratory (Max Delbrück Center) to access mouse samples. The group is currently working in DNA repair and CSR (*69*, *158*, *159*) using extensively CH12-F3 cells (CH12), a mouse lymphoma B cell line that almost exclusively switches to IgA (*76*). They kindly provided me with wild-type (WT) and clustered regularly interspaced short palindromic repeats / CRISPR associated protein 9 (CRISPR/Cas9) knockout CH12 (Rif1$^{-/-}$, Ligase 4$^{-/-}$, 53BP1$^{-/-}$ and BRCA1$^{-/-}$) (Table 2) that I activated *in vitro* using IL4, mouse transforming growth factor β1 (mTGF-β1) and anti-CD40L. Also, they provided me with WT mouse intestines from which I isolated primary mouse B cells from Peyer's patches (PP) that have switched *in vivo* mainly to IgA (*160*).

The mouse switch-joint PCR was designed following the human workflow, which only allows the amplification of switch regions in IGH loci subjected to class-switching. I designed primers based on the murine switch region annotations (Table 8) and, in the case of mSM FW primer, based on Shinkura et al., 2004 (*38*) (Figure 11A). Switch-joint PCR worked for all the designed primers in *in vitro* and *in vivo* activated mouse B cells (0BC); however, when the FW and RV primers carried 20 nt barcodes (2BC), all amplifications decreased their efficiency (Figure 11B). Thus, I decided to amplify the mouse switch-joints combining the barcoded mSM FW primer with an unbarcoded RV primer (1BC), which did not limit the efficiency of the PCR to such a significant extent (Figure 11B). PP and CH12 contained mainly SM-SA switch-joints and SM-SG2bc. PP contained many SM-SG1 joints, unlike CH12, whose third most frequent joint was SM-SG3. Analysis was done in CH12 using only the mSM FW primer and the mouse SA (mSA), mouse SG3 (mSG3) and mouse SG2b and SG2c (mSG2bc) RV primers, excluding mouse SG1 RV primer. The reason was that CH12 elicited less SM-SG1 than SM-SA, SM-SG2bc or SM-SG3. The translation to murine samples in the SWIBRID pipeline involved the usage of the genome assembly GRCm38 and the annotations of the murine switch regions (Table 8).

I tested the newly translated SWIBRID using WT *in vivo* and *in vitro* activated mouse samples. The *in vivo* activated sample, PP, elicited 395 eff_nclusters, of which 4.3% were SG3, 30.3% were SG2b, 3.5% were SG2c, and 59.7% were SA. On the other hand, the *in vitro* activated sample, CH12, showed 417 eff_nclusters, of which 90.4% were SA. Most clones were derived from SM-SA class-switching, which is expected due to the nature of PP and CH12 (*76*, *160*). Regarding diversity, the four most prominent clusters in CH12 occupied 47.3% of all the reads, while in PP, only 6.9%. This result showed that CH12 elicited a higher diversity than PP, although the cluster size distribution was more conserved in PP (Figure 11C).

In conclusion, the pipeline was successfully translated to mouse samples. We found a lower isotype diversity in mouse samples compared to human samples, which is expected due to the nature of the cell line, CH12, and the origin of the mouse I obtained PP from, a non-immunized mouse.

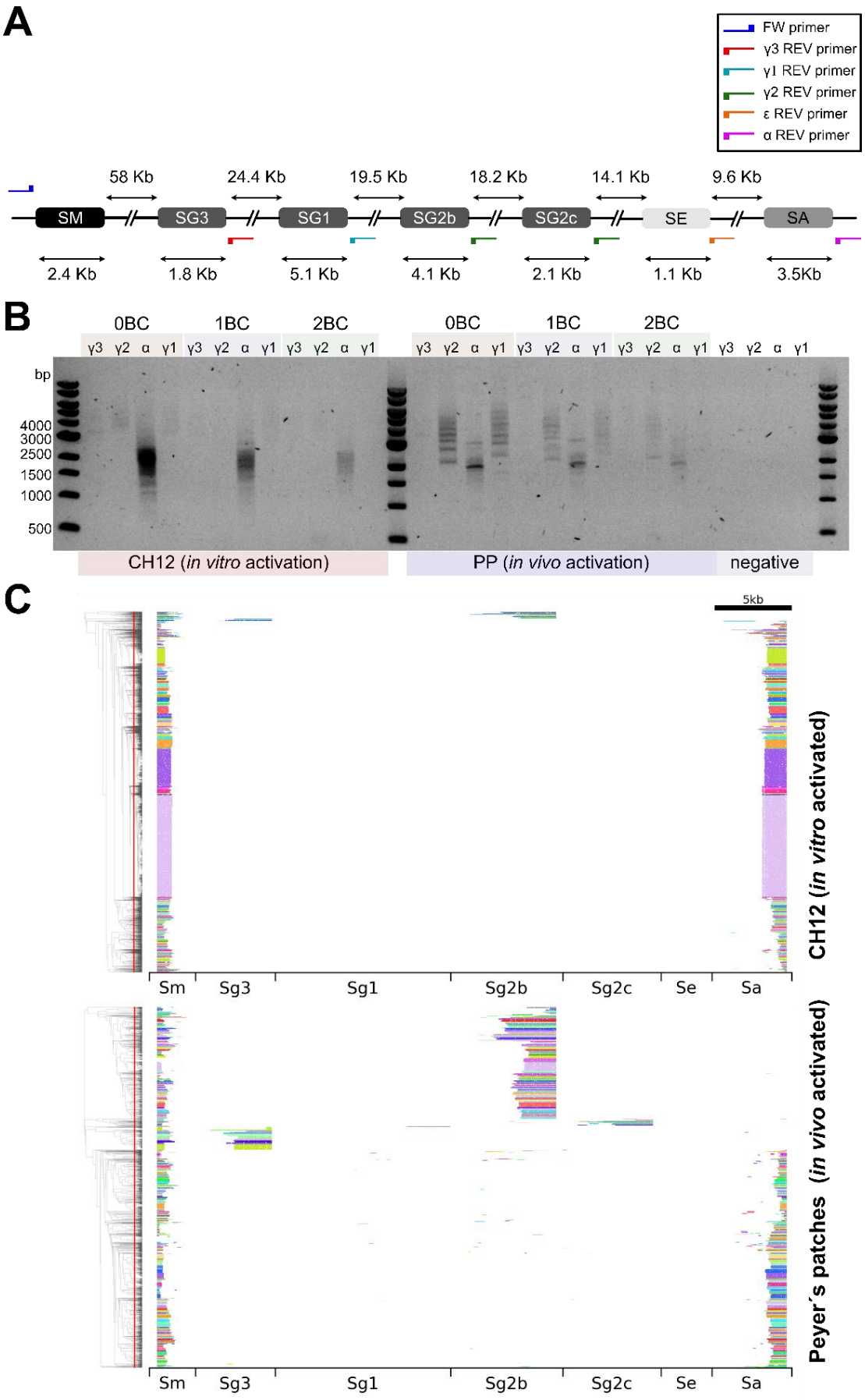


Figure 11. Translation of SWIBRID for the analysis of mouse samples

A. Switch regions of the mouse IGH locus show primer binding sites in different colors. Kb=kilobases. SM=switch μ region (GenBank: AF446347.1); SG3=switch γ3 region (GenBank: D78343.1); SG1=switch γ1 region (GenBank: M12389.2); SA=switch α region (GenBank: D11468.1); SG2b=switch γ2b region (*142*); SG2c=switch γ2c region (*142*) and SE=switch ε region (GenBank: M17012.1). **B.** Amplicons of switch-joints from μ to γ3, μ to γ2bc, μ to α and μ to γ1 in a 0.8% agarose gel. The material used was activated CH12 (*in vitro*) and activated mouse Peyer's patches (PP) (*in vivo*). 0BC=PCR with unbarcoded primers, 1BC=PCR using FW primer barcoded and RV primer unbarcoded, 2BC=PCR using FW and RV barcoded primers. **C.** Readplots of *in vivo* and *in vitro* activated mouse cells (Peyer's patches and CH12, respectively). 5 000 reads are shown in the plots. On the left is a hierarchical clustering dendrogram of the switch-joint reads and the cut-off (red) used to determine the number of clusters. Each row represents an individual switch-joint read, and the x-axis represents the switch regions' coordinates. Reads belonging to the same cluster are depicted in the same color.

7.3. Analysis of proteins involved in the insert incorporation mechanism unravels SWIBRID limitations

So far, the mechanism of J-CH1 insert acquisition in the switch regions remains elusive. As we developed a tool capable of identifying J-CH1 inserts and individual B cell clones in a sample, I am now positioned to address the molecular players of the insert incorporation. The insert incorporation mechanism can be divided into i) insert source, ii) switch region donor break inducer and iii) insert repair into the switch regions. However, first, we needed to decide if the analysis would be performed in human or mouse samples.

7.3.1. Insert frequency is stable within different replicates of the same donor in humans

Analysis of insert frequency was previously standardized using reads (*94*). However, sample reads do not resemble sample diversity, making the use of reads incorrectly to normalize the analysis of inserts. Considering the usage of switch-joint B cell clones as a novel way of standardizing the insert frequency, I decided to test the natural J-CH1 insert acquisition *in vivo* in different human HD.

Six HD were analyzed using SWIBRID to calculate the unique J-CH1 inserts per eff_nclusters (J-CH1 insert frequency). The J-CH1 insert frequency within the same donor was consistent regardless of the cell number input (Figure *12*A). However, the J-CH1 insert frequency in all six HD at cell numbers lower than 50 000 was irregular, indicating that the analysis of J-CH1 insert frequency requires at least 50 000 input cells to provide a reliable result. This issue reveals that SWIBRID has limitations that could lead to false results when analyzing samples deficient in B cell clones. Despite the J-CH1 insert frequency being consistent within the same donor, the HD had completely different J-CH1 insert frequencies suggesting that every donor has a unique signature (median J-CH1 insert frequency per 1 000 eff_nclusters: HD1=13.4, HD2=8.6, HD3=15.2, HD4=5.9, HD5=6.5 and HD6=4.9) (Figure *12*B).

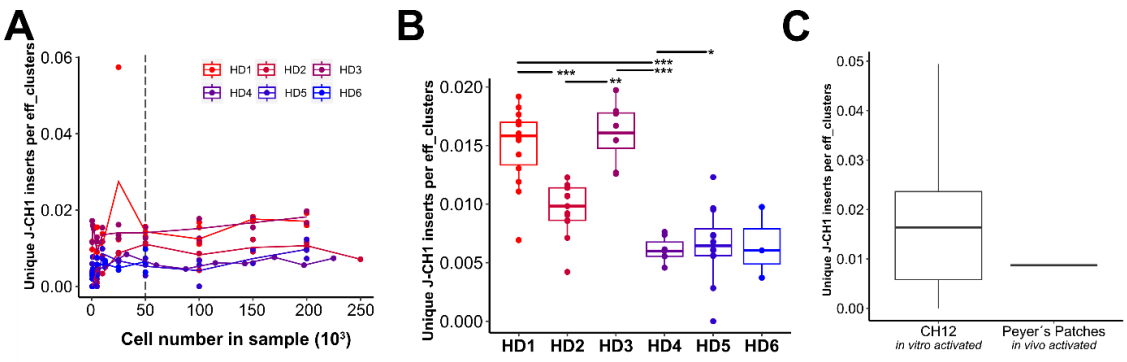


Figure 12. Insert frequency varies between healthy human donors.

A. Unique J-CH1 inserts per eff_ncluster in six HD were analyzed using different amounts of cells. Three biological replicates were performed in HD1, HD2 and HD3; two biological replicates were performed in HD4 and only one replicate was performed in HD5 and HD6. HD=healthy donor. HD1, HD2, HD3, HD4 – memory IgGA B cells; HD5, HD6 – total B cells. Every HD is represented by a different color. **B.** Unique J-CH1 insert per eff_ncluster is shown for different HD. The biological replicates samples used contained more than 49 000 cells. $HD1_n$=14, $HD2_n$=13, $HD3_n$=8, $HD4_n$=7, $HD5_n$=12, $HD6_n$=3. Normality was found in all samples, and a ANOVA post-hoc test was performed. P-value <0.05=*, p-value < 0.01=** and p-value < 0.001=***. **C.** Unique J-CH1 insert per eff_ncluster is shown in murine samples. WT variants of CH12 (n=8) and Peyer's patches (n=1).

In vivo and *in vitro* activated mouse B cells (CH12 and Peyer's patches (PP)) were analyzed by SWIBRID. J-CH1 inserts were identified in both samples (mean unique J-CH1 inserts in CH12: 14.3, PP: 9) (Figure *12*C). Direct comparison of *in vivo* J-CH1 insert acquisition in human and mouse samples showed that the latter elicited slightly lower J-CH1 insert frequency than human samples (PP=8.7 versus HD=9.2 unique J-CH1 inserts per 1 000 eff_nclusters).

In conclusion, we found that the J-CH1 insert frequency is stable within the same HD as long as the analysis is conducted on a minimum of 50 000 cells. Interestingly, the HD had different J-CH1 insert frequencies, suggesting specific J-CH1 insert frequencies for different donors potentially related to their DNA repair machinery. Even though the J-CH1 insert frequency was barely the same, I chose human samples to analyze the mechanism of the J-CH1 inserts because they elicit more eff_nclusters than mouse samples, making the probability of finding a J-CH1 insert higher.

7.3.2. Study of J-CH1 insert acquisition by the modulation of DNA repair

Lebedin et al. 2022, studied the characteristics of J-CH1 inserts in an attempt to understand their incorporation mechanism; however, J-CH1 insert acquisition studies *per se* have not yet been performed (*97*). The study of J-CH1 insert acquisition comprises the modulation of proteins *in vitro* to learn protein

contributions to that event. It is hypothesized that J-CH1 inserts in switch regions need a broken acceptor site, which has been granted to the activation-induced cytidine deaminase (AID) since memory B cells contain a higher amount of J-CH1 inserts than naïve B cells (*97*). Therefore, the J-CH1 insert acquisition studies were performed in a model of activation of naïve B cells ($CD27^{-}IgG^{-}IgA^{-}IgM^{+}IgD^{+}$) under the treatment of DNA repair modulating agents (Olaparib, SCR7). J-CH1 inserts need a DNA donor source, which has been suggested to be genomic loci close to early replicating fragile sites (ERFS), as studied in Lebedin et al. 2022 (*97*). Thus, we decided to provoke ERFS instability using hydroxyurea (HU) with the aim of increasing the J-CH1 insert DNA source (*161*, *162*). Finally, J-CH1 inserts incorporation into the switch regions is likely to finish by ligating the DNA fragment into the IGH. Considering that c-NHEJ and a-EJ are the preferred DNA repair mechanisms during CSR, we decided to use SCR7 to impair c-NHEJ (*163*) and Olaparib to impair a-EJ (*164*) and see their effect in the J-CH1 insert acquisition (Figure *13*A).

In vivo and *in vitro* activated B cells were analyzed. *In vivo* activated B cells were isolated from blood as $CD27^{+}IgG^{+}IgA^{+}$ B cells (Memory), while *in vitro* activated B cells derived from naïve B cells isolated from the blood and activated in culture using IL-4 and CD40L (Activated). Activated and Memory B cells were analyzed to be compared to each other. On the other hand, B cells were also activated *in vitro* and treated with 80 μM SCR7, 1.25 μM Olaparib, 25 μM HU and DMSO (as control, since it is the solute for the DNA repair modulators) to see their effect on J-CH1 insert acquisition (Figure *13*B). Concentrations were chosen after chemical titration.

SWIBRID analysis identified the highest number of eff_nclusters in Activated and HU samples with values higher than 2 500 eff_nclusters. Following previous results that indicated the need for a minimum input of 50 000 cells in the analysis (Figure *12*A), I decided to establish a minimum amount of eff_nclusters to be considered a sample suitable for analysis. Taking into account that the PCR in the J-CH1 insert acquisition experiments was not optimized to the number of cells but to the amount of gDNA, I considered correct discarding samples with less than 100 eff_nclusters (Figure *13*C). Despite having less diversity, Memory B cells had a higher number of unique J-CH1 inserts than Activated, while for treated samples, SCR7 samples elicited a higher number of unique J-CH1 inserts (average unique J-CH1 inserts in Memory=17.3±10.9 and SCR7=22). Unfortunately, most SCR7-treated samples did not elicit more than 100 eff_nclusters impeding the statistical comparison of the data with other samples. The lowest unique J-CH1 inserts were found in Olaparib-treated samples (average unique J-CH1 inserts=2±1.41). Differences in unique J-CH1 inserts were not found among the rest of the samples (average unique J-CH1 inserts in Activated=15.75±14.6, HU=17.33±15, DMSO=12±15.5) (Figure *13*D).

Comparison of unique J-CH1 inserts per eff_nclusters (J-CH1 insert frequency) did not reveal a big difference between Activated and Memory (median J-CH1 insert frequency J-CH1 insert frequency in Activated=0.0032±0.003 and Memory=0.004±0.002). However, there was a tendency that indicated a lower J-

CH1 insert frequency in samples treated with Olaparib (median J-CH1 insert frequency in Olaparib=0.0005±0.001, HU=0.0013±0.025, SCR7=0.0047±0.006, DMSO=0.0020±0.003) (Figure *13*E). High variance among the biological replicates within the same conditions indicates that the J-CH1 insert frequency potentially varied among the different donors, as we have seen in previous results (Figure *12*B). It could also be explained by the fact that these donors had a silent DNA repair deficiency, which could only be addressed by exome sequencing. Addressing this issue should involve using chemicals at different concentrations within the same donor or having the same donor represented by more than one replicate.

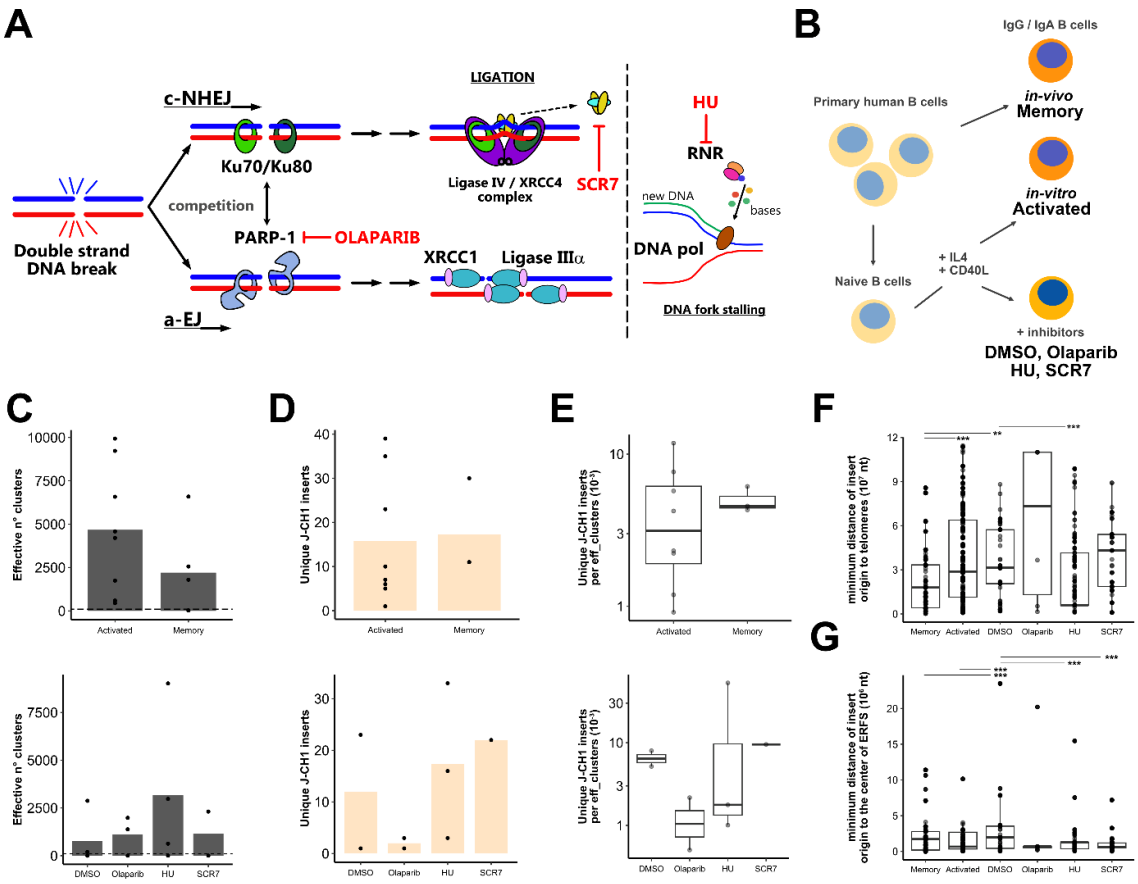


Figure 13. Modulating DNA repair does not elicit differences in J-CH1 insert frequencies but affects their origin.

A. Scheme depicting the action point of DNA repair inhibitor SCR7 and Olaparib in DNA repair pathways and the DNA replication fork stall chemical, hydroxyurea (HU). PARP-1=Poly [ADP-ribose] polymerase 1, X-ray repair cross-complementing protein 1 (XRCC1), RNR=ribonucleotide reductase, DNA pol=DNA polymerase, cNHEJ=classical non-homologous end joining, aEJ=alternative end joining. DNA (+)-sense in blue, (-)-sense in red and newly produced DNA in green. A straight red line with a perpendicular end indicates inhibition. **B.** Scheme depicting isolation and activation of primary human B cell populations. IL4=interleukin 4 and CD40L=Cluster of differentiation 40 ligand. **C.** The eff_ncluster of Activated versus Memory B cells (top) and after treatment

with different inhibitors (bottom). Bar plots show the mean of the effective clusters. The dashed line indicates 100 eff_nclusters indicating the limit for analysis of J-CH1 insert frequency. **D.** Unique inserts found in activated versus memory B cells (up) and after treatment with different inhibitors (below). Bar plots show the mean of the J-CH1 inserts in samples with more than 100 eff_nclusters. **E.** J-CH1 insert frequency in Activated versus Memory B cells (up) and after treatment with different inhibitors (below) in samples with more than 100 eff_nclusters. The boxplot line represents the 50% quartile. Biological replicates in Activated=4; Memory=5; DMSO=5; Olaparib=4; HU=5; SCR7=3. **F.** Median distance (med.dist) of J-CH1 insert origins to the telomeres. nt=nucleotides. The boxplot line represents the 50% quartile, the upper box the 75% percentile and the lower the 25% percentile. **G.** Median distance (med.dist) of J-CH1 insert origin to the closest early replication fragile sites (ERFS). The boxplot line represents the 50% quartile, the upper box the 75% percentile and the lower the 25% percentile. Inserts in Activated=543; Memory=775; DMSO=111; Olaparib=6; HU=290; SCR7=103. Shapiro test was performed, but no normal distribution was found in the samples. Pair-wise Wilcox rank sum test was performed to test the difference between samples. P-value <0.05=*, p-value < 0.01=** and p-value < 0.001=***.

As previously mentioned, it was identified that J-CH1 insert origin in naïve B cells was closer to telomeres, while J-CH1 insert origin in activated B cells was not only closer to telomeres but especially to ERFS (*97*). Interestingly my results show that the origin of J-CH1 inserts acquired in Memory and Activated samples had a similar distance to ERFS, but Memory J-CH1 insert origins were closer to telomeres (median distance (med.dist) to telomeres in Memory=17 993 157 nt and Activated=28 811 293 nt). Treatment with HU provoked a shift in J-CH1 insert origin that located them closer to telomeres and ERFS in comparison to DMSO (med.dist telomeres in DMSO=31 380 685 nt and HU=5 965 255 nt; med.dist ERFS in DMSO=1 975 431 nt and HU=1 262 147 nt). Inhibition of Ligase 4 by SCR7 also shifted the J-CH1 insert origin closer to ERFS (med.dist ERFS=643 671.5 nt). Olaparib-treated samples only acquired six J-CH1 inserts that showed high variability in med.dist ERFS and med.dist telomeres. The high variability among the few J-CH1 insert origins suggests that PARP1, inhibited by Olaparib, could be involved in J-CH1 inserts acquisition (Figure *13*F-G).

In summary, *in vivo* and *in vitro* activated B cells (Memory and Activated) showed a similar J-CH1 insert acquisition, while cells activated under Olaparib treatment showed a lower J-CH1 insert acquisition. Analysis of J-CH1 insert origin revealed that J-CH1 inserts acquired in cells activated under HU treatment were closer to telomeres and ERFS and that *in vivo* activation of B cells shifted J-CH1 insert origin to be closer to telomeres.

7.3.3. Inhibition of ATM impairs J-CH1 insert acquisition and B cell diversification

The DNA inhibition of Ligase 4 and PARP1 and the replication fork stall in different donors did not elicit conclusive results. The lack of technical replicates in treated samples could explain the inconclusive results in conditions like SCR7

treatment that only elicited one sample for analysis. In order to contemplate the impact that individual J-CH1 insert frequencies could have on the results of our experiments, I decided to focus on studying a single protein.

Ataxia-telangiectasia mutated (ATM) was chosen to be studied for its implication in J-CH1 insert acquisition. ATM is one DNA double-strand break-sensing protein that starts the cascade of DNA repair (*165*). We decided to study ATM involvement in J-CH1 insert acquisition because if the repair of dsDNA breaks were involved in ligating DNA fragments to the IGH to become J-CH1 inserts, we would impair their acquisition by inhibiting ATM since ATM is a dsDNA break sensing protein.

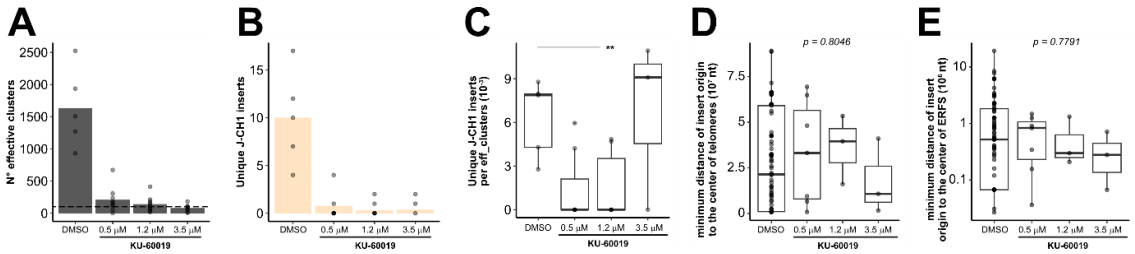


Figure 14. ATM inhibition impairs B cell diversification and does not affect J-CH1 insert origin.

A. The eff_nclusters of primary human B cells activated under the treatment of KU-60019 or DMSO (control). The dashed line indicates 100 eff_nclusters. **B**. Number of unique J-CH1 inserts of primary human B cells activated under the treatment of KU-60019 or DMSO that contains more than 100 eff_nclusters. **C.** J-CH1 insert frequency of primary human B cells activated under the treatment of KU-60019 or DMSO that contains more than 100 eff_nclusters. Boxplot-line shows the 50% quartile. DMSO: 2 tech. rep and 3 donors; KU-60019 0.5 µM: 2 tech. rep and 3 donors; KU-60019 1.2 µM: 2 tech. rep and 3 donors and KU-60019 3.5 µM: 2 tech. rep and 3 donors **D.** Median distance (med.dist) of J-CH1 insert origins to the telomeres, nt=nucleotides. The boxplot line represents the 50% quartile, the upper box the 75% percentile and the lower the 25% percentile. **E.** Median distance (med.dist) of J-CH1 insert origins to ERFS, nt=nucleotides. The boxplot line represents the 50% quartile, the upper box the 75% percentile and the lower the 25% percentile. The number of J-CH1 inserts in DMSO: 117; KU-60019 0.5 µM: 7; KU-60019 1.2 µM: 3 and KU-60019 3.5 µM: 3.. Shapiro test was performed, and not all the samples were not normal; thus, the Kruskal-Willis test was performed to compare the samples. **=p-value < 0.01.

Activation of primary human naïve B cells under the treatment of ATM inhibitor (KU-60019) was performed at different concentrations (0.5 µM, 1.2 µM and 3.5 µM). Very low concentrations of KU-60019 were enough to disrupt B cell diversification significantly in three donors. Considering that inhibition of ATM is inhibiting dsDNA break sensing and that activated B cells undergo CSR, which involves dsDNA breaks, we could expect a decrease in the number of

eff_nclusters due to the apoptosis of cells in which DNA breaks were left unrepaired. Control-activated samples (DMSO) elicited a mean eff_nclusters of 1 633, while the mean eff_nclusters of the samples treated with KU-60019 were >7-fold lower. (mean eff_nclusters at 0.5 µM=208, 1.2 µM=139 and 3.5 µM=78.5 of KU-60019) (Figure *14*A). We discarded samples with less than 100 eff_nclusters and analyzed the unique J-CH1 inserts and their frequency. KU-60019 treatment greatly disrupted J-CH1 insert acquisition, considering that samples elicited a mean of less than 5 unique J-CH1 inserts, while DMSO elicited a mean of 10 unique J-CH1 inserts (Figure *14*B). As expected by the results, J-CH1 insert frequency was lower for the samples treated with low KU-60019 compared to DMSO, and specifically, 1.2 µM KU-60019 J-CH1 insert frequency was significantly lower than DMSO (Wilcox test, p-value=0.006464). Controversially, the three samples carrying more than 100 eff_nclusters treated with 3.5 µM KU-60019 elicited a normal J-CH1 insert frequency (Figure *14*C). No significant difference was found in the J-CH1 insert origin proximity to telomeres or ERFS**,** which might be explained by the general lack of J-CH1 inserts acquired under the KU-60019 treatment (Figure *14*D-E).

In conclusion, ATM inhibition disrupts J-CH1 insert acquisition at low KU-60019 treatment and B cell diversification. Interestingly, the J-CH1 insert origin under the treatment of KU-60019 was unchanged. Unfortunately, many samples were discarded from the analysis due to poor B cell diversification.

7.3.4. Telomerase inhibition particularly affects J-CH1 insert acquisition in switch mu region

Telomeres are DNA regions with well-orchestrated DNA machinery regulating replication and repair. One of the most special proteins in the maintenance of telomeres is telomerase reverse transcriptase (TERT), one of the few reverse transcriptase RNA polymerase in human cells (*166*). Considering that J-CH1 insert origin in naïve B cells is closer to telomeres (*97*) and that expression of TERT is increased upon B cell activation (*122*, *123*), we believe that TERT could contribute to the generation of J-CH1 insert substrates. BIBR1532 is a chemical that inhibits the repetitive sequence addition of TERT (*167*). Thus, I decided to inhibit TERT function using BIBR1532 and screen J-CH1 insert frequency and characteristics at different concentrations.

Considering that 10 µM of BIBR1532 maintains myeloma cell lines alive, but the double amount lowers the viability to 60% (*168*, *169*), we decided to start the treatment of primary human B cells using up to 5 µM BIBR1532. Naïve B cells from three donors were isolated and activated under the treatment of BIBR1532 (1 µM, 2 µM, 5 µM) or DMSO as a control. A very similar amount of eff_nclusters was found among all the samples, discarding only one sample with less than 100 eff_nclusters for further analysis (1 µM BIBR1532 of HD7 with 88 eff_nclusters) (Figure *15*A). The highest mean quantification of unique J-CH1 insert was found in BIBR1532 1 µM, but it was not significantly different (Figure *15*B). J-CH1 insert frequency was also similar for samples treated with different concentrations of BIBR1532 and DMSO (J-CH1 inserts per 1 000 eff_nclusters, DMSO=4.1, 1 µM=4.3, 2 µM=4.5, 5 µM=4.3 of BIBR1532) (Figure *15*C). Analysis of J-CH1

inserts origin showed that high concentrations of BIBR1532 significantly shifted J-CH1 insert origin closer to ERFS. In contrast and unexpectedly, J-CH1 insert origin distance to telomeres was maintained across all samples, except for BIBR1532 1 µM, whose origin of J-CH1 inserts showed significantly closer proximity to telomeres than DMSO control (Wilcox's test, p-value=0.0002498) (Figure *15*D-E). Considering that only BIBR1532 1 µM showed such a result, I cannot conclude that inhibition of TERT increases the J-CH1 insert origin closer to the telomeres.

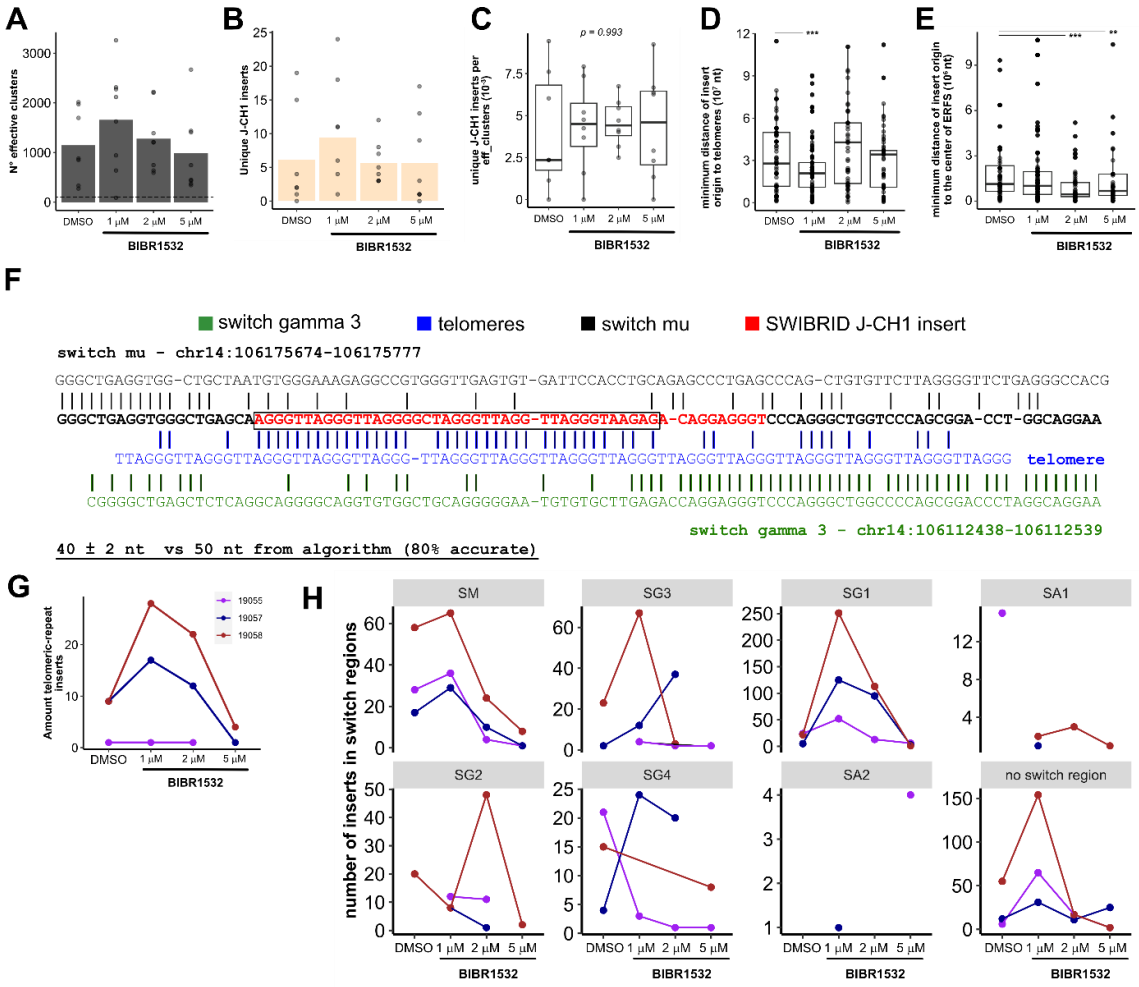


Figure 15. TERT inhibition during B cell activation impairs in a dose-response manner the ability to acquire J-CH1 inserts within SM, but it does not affect the total J-CH1 insert frequency.

A. The eff_nclusters of primary human B cells activated in the presence of indicated concentrations of BIBR1532 or DMSO as a control. The dashed line indicates 100 eff_nclusters. **B**. Number of J-CH1 inserts of primary human B cells activated under the treatment of BIBR1532 or DMSO as a control. **C.** J-CH1 insert frequency of primary human B cells activated under the treatment of BIBR1532 or DMSO as a control. Boxplot-line shows the median. ANOVA was performed. **D.** Median distance (med.dist) of J-CH1 insert origins to the telomeres under the treatment of BIBR1532 or DMSO as a control.

nt=nucleotides. Normality was not found in these samples and a wilcox-test was performed to compare the groups. **E.** Median distance (med.dist) of J-CH1 insert origins to the ERFS under the treatment of BIBR1532 or DMSO as a control. nt=nucleotides J-CH1 inserts originating in telomeres or chrUn (pseudo-chromosomes) were not considered in D and E. In C-E, the boxplot line represents the 50% quartile, with the respective 75% percentile and the 25% percentile indicated by the upper and lower bars. Normality was not found in these samples and a wilcox-test was performed to compare the groups. Inserts found in the samples were DMSO=131, BIBR1532 1 µM: 256, BIBR1532 2 µM: 58 and BIBR1532 5 µM: 70. **F.** Manual analysis of one representative read containing a "telomeric repeat". Straight lines connect matching nucleotides of two different sequences. In bold can be found the raw sequence of the read. Highlighted in red is the part of the read identified as a "telomeric insert" by SWIBRID. The black box highlights the telomeric repeat. Genomic coordinates are written following GRCh37. Telomeric repeats were found in DMSO=19, BIBR1532 1 µM=46, BIBR1532 2 µM=35 and BIBR1532 5 µM=5. Total manually analyzed reads=10. **G.** Quantification of telomeric-repeats found as J-CH1 insert in the presence of different concentrations of BIBR1532 upon activation. **H.** Count of J-CH1 inserts within different switch regions depending on the concentration of BIBR1532 used upon activation. Colors represent donors (the legend can be found in G). For A-E, G and H: DMSO=2 tech. rep and 3 donors; BIBR1532 1 µM=2 tech. rep and 3 donors; BIBR1532 2 µM=2 tech.rep and 3 donors and BIBR1532 5 µM=2 tech.rep and 3 donors. The Shapiro test was performed to test the normality of the samples. A t-test was performed in C. In contrast, the Wilcox test was performed in D and E. *=p-value < 0.05.

The inhibition of TERT impairs the elongation of the telomeres, which is necessary for the survival of highly proliferative cells such as activated B cells. The elongation of telomeres is a process that involves the reverse transcription of telomeric repeats. This process is also thought to occur in non-telomere areas of the genome since interstitial telomeric sequences (ITS) can be found in the genome (*130*). Following the hypothesis that telomeric repeats could form J-CH1 inserts, we found at least ten in every condition. Since telomeric repeats and switch regions are repetitive, I manually analyzed ten random reads containing these J-CH1 inserts and confirmed their veracity (Figure *15*F). J-CH1 inserts that are telomeric repeats decreased in a dose-dependent manner in two out of three donors. One of the donors did not elicit any telomeric repeat (Figure *15*G). The lack of telomeric repeat J-CH1 insert could be analyzed together with TERT activity in the samples. The decrease is not significantly different from DMSO but shows a dependency on TERT for generating such a J-CH1 insert source.

On another note, Lebedin et al. 2022 found that the origins of J-CH1 inserts from naïve B cells were the closest to the telomeres (*97*). Naïve B cells do not go through CSR; thus, the J-CH1 inserts found in naïve B cells must have been inserted between the J and the IGHM domain, likely in the SM. In addition, naïve B cells highly express TERT to overcome the proliferation burden that B cell

maturation entails. Considering these two facts, I analyzed how many J-CH1 inserts were acquired within each switch region after inhibiting TERT. Despite no difference in J-CH1 insert acquisition within SGs and SAs, J-CH1 inserts acquired within SM decreased upon BIBR1532 treatment in a dose-dependent manner (Figure *15*H). The data suggest that TERT could have a role in the J-CH1 insert acquisition in SM.

In summary, inhibition of TERT did not impair the general J-CH1 insert acquisition. However, it disrupted telomeric repeat J-CH1 insert acquisition. Also, inhibition of TERT impaired the J-CH1 insert incorporation within SM, suggesting that TERT could play a role in J-CH1 insert acquisition in that switch region.

7.3.5. AID$^{-/-}$, ATM$^{-/-}$, BRCA1$^{-/-}$ and Ligase 4$^{-/-}$ patients show different J-CH1 insert acquisition than HD

Modulating DNA repair by applying modulators *in vitro* may not reflect human protein deficiencies because impairment of protein function is not equal to the absence of the protein. In fact, the mere presence of proteins, e.g., dead-kinase ATM, is sufficient to carry out some DNA repair functions (*170*). Hence, we established a collaboration with Dr. Qiang Pan-Hammarström (Karolinska Institute, Sweden) to analyze J-CH1 insert frequency in DNA repair-deficient patients (DRdef) carrying deficiencies in AID, ATM, breast cancer gene 1 (BRCA1), Ligase 4 and nipped-B-like protein (NIPBL).

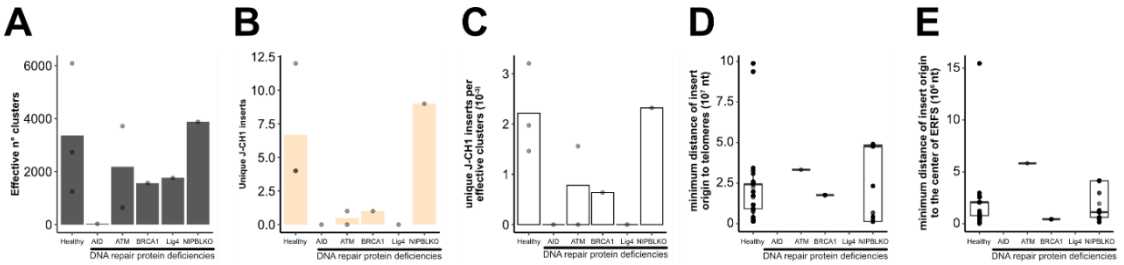


Figure 16. Natural J-CH1 insert acquisition is deficient in BRCA1$^{-/-}$, ATM$^{-/-}$, AID$^{-/-}$ and Ligase 4$^{-/-}$ patients.

A. The eff_nclusters of DNA repair-deficient patients (DRdef) compared to HD. **B**. Number of J-CH1 inserts of DRdef compared to HD. **C.** Insert frequency of DRdef samples compared to HD. **D.** Median distance (med.dist) of J-CH1 insert origins to the telomeres, nt=nucleotides. **E.** Median distance (med.dist) of J-CH1 insert origins to the ERFS, nt=nucleotides. J-CH1 inserts originating in telomeres or chrUn (pseudo-chromosomes) were not considered in D and E. Donors analyzed for: Healthy, n=3; AID, n=1; ATM, n=2; BRCA1, n=1, Ligase 4, n=1 and NIPBL, n=1. Unique J-CH1 inserts (i) in Healthy, i=21; AID, i=0; ATM, i=1; BRCA1, i=1, Ligase 4, i=0 and NIPBL, i=9. In A-C the barplots represent the mean. In D-E, the boxplot line represents the 50% quartile, with the respective 75% and 25% percentile indicated by the upper and lower lines of the box.

AID is the promoting enzyme of CSR; without it, CSR is not expected to occur. Thus, analysis of DRdef showed that, $AID^{-/-}$ did not elicit a high number of eff_nclusters ($AID^{-/-}$=30). Thus, $AID^{-/-}$ was not included in further analysis (Figure *16*A). We did not find J-CH1 inserts in deficiencies of $ATM^{-/-}$, Ligase $4^{-/-}$ and $BRCA1^{-/-}$, unlike $NIPBL^{-/-}$, which elicited 9 unique J-CH1 inserts, higher than 6.6, the average unique J-CH1 insert found in HD (Figure *16*B). Nonetheless, $NIPBL^{-/-}$ J-CH1 insert frequency was in the same range as in HD ($NIPBL^{-/-}$=0.0023 and HD=0.0022) (Figure *16*C). Analysis of J-CH1 insert origin showed similar proximity to telomeres and ERFS of J-CH1 inserts acquired in HD and $NIPBL^{-/-}$ (Figure *16*D-E).

In conclusion, and considering that these samples comprise single case studies due to the rarity of the disease, the data suggest that $BRCA1^{-/-}$, $ATM^{-/-}$, $AID^{-/-}$ and Ligase $4^{-/-}$ impair J-CH1 inserts, while $NIPBL^{-/-}$ does not do so when compared to HD.

7.3.6. Characteristics of SWIBRID J-CH1 inserts in HD agree with published RNA J-CH1 insert data

Lebedin et al., 2022 analyzed inserts in antibody transcripts of HD. They discovered that inserts could be classified by their incorporation position and the B cell subpopulation where they were found (*97*). SWIBRID inserts represent insertions in the genomic switch region, and they would classify as J-CH1 inserts since the switch region is located between the J segment and CH1 exon. I decided to analyze the characteristics of the J-CH1 inserts found in HD and follow the analysis according to Lebedin et al. 2022.

Firstly, I decided to compare the origin of the J-CH1 inserts found by SWIBRID in DNA to those found in transcripts, specifically if they derive from exonic, intronic or intergenic regions. We detected 31 880 J-CH1 inserts in HD with unique origins. To ease the analysis, I collected 3 sets of 500 unique J-CH1 inserts randomly chosen and ensured that J-CH1 insert origin did not overlap among the sets. As a control, I generated 3 *in silico* datasets following the same lines as Lebedin et al., 2022. The control data must be considered to derive from genic or intergenic regions. Thus, the control *in silico* insert datasets were randomly originated from exonic, intronic and intergenic regions in the exact percentage as their counterpart set of 500 J-CH1 inserts. The randomization was done within the same chromosome and at the same length as the original J-CH1 inserts (Figure 17A).

SWIBRID origins of SWIBRID J-CH1 inserts were identified as 6.4±0.8% exons, 62±1.3% introns and 31.5±2.1% intergenic. Comparison of SWIBRID J-CH1 inserts to Lebedin et al. 2022 inserts showed that SWIBRID inserts were more similar to the RNA V-DJ inserts (exon=29.1%, intron=42.1%, intergenic=28.8%) than to the RNA J-CH1 inserts (exon=69.4%, intron=18.45%, intergenic=12.15%) (Figure 17B) (*97*). To exclude that the J-CH1 inserts source is randomly generated from the genome, I compared them to a randomized simulated dataset without considering their genic origin. Therefore, J-CH1 inserts were generated randomly, only considering their chromosome and length. Comparison of simulated J-CH1 inserts to SWIBRID J-CH1 inserts showed that, unlike SWIBRID

J-CH1 inserts, simulated inserts were enriched in intergenic regions (exon=1.6%, intron=42.8%, intergenic=55.6%). In an attempt to find common features from SWIBRID J-CH1 inserts donor origins, DNA methylation frequency and median distance to the closest R-loop, ERFS and telomeres were measured.

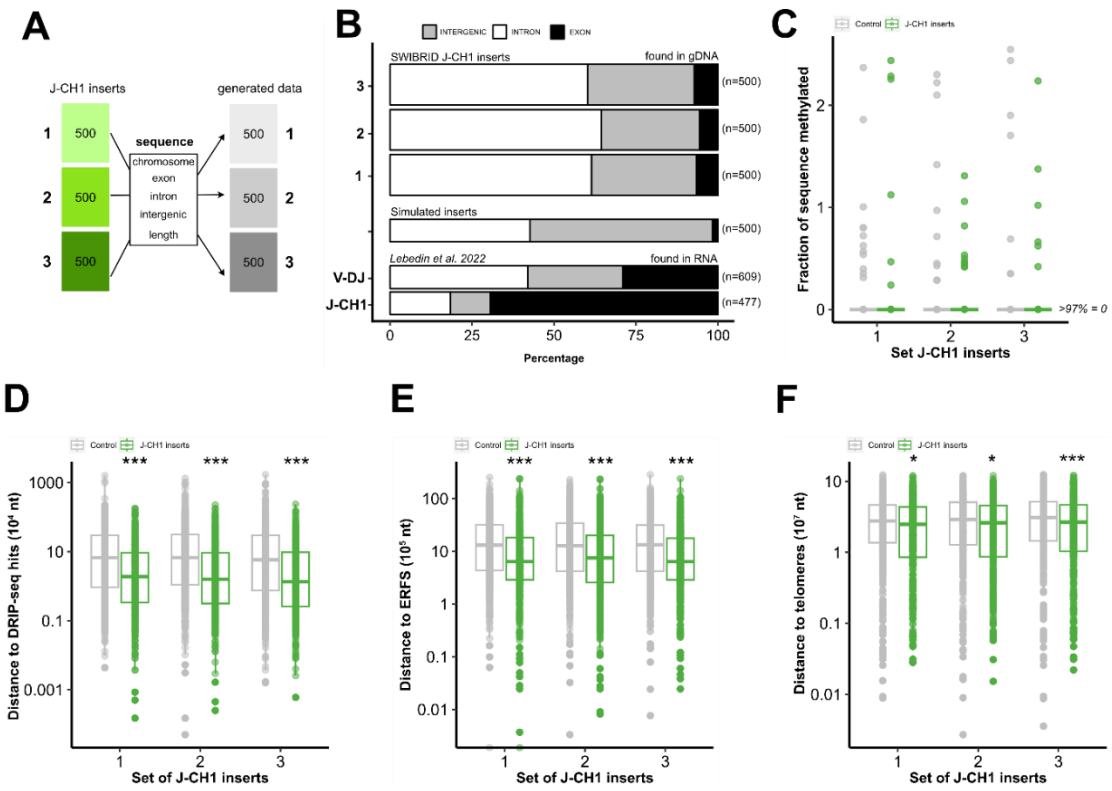


Figure 17. J-CH1 insert characteristics resemble V-DJ inserts characteristics found in RNA.

A. Scheme of control data generation to analyze the J-CH1 inserts. **B.** Classification of inserts by origin from genes: introns (white), intergenic (grey) or exons (black) regions. Data of V-DJ and J-CH1 inserts found in RNA were obtained from Lebedin et al., 2022 (*97*). **C.** Fraction of J-CH1 insert that originates from methylated DNA. Data were obtained from methylation studies during transcription upon B cell activation (*172*). **D-F.** Median distance (nt) of the J-CH1 insert original site to the closest R-loop (determined by DRIP-seq) (*173*) (**D**), early replicating fragile sites (ERFS) (*162*) (**E**) and telomeres (**F**). In D-F, the boxplot line represents the 50% quartile and the upper and lower lines of the box the respective 75% and 25% percentile. The Wilcox test was performed. p-value <0.05=*, <0.001=***.

DNA methylation is a DNA modification that shapes the epigenetics in the genome. It has been observed that DNA hypomethylation can lead to genomic instability (*171*). Therefore, I quantified the DNA methylation frequency in the J-CH1 insert origin using a methylation dataset obtained from a B cell differentiation analysis (*172*). 97% SWIBRID J-CH1 insert donor origin did not coincide with methylated positions; thus, they can be considered hypomethylated. However,

there was no difference between the control and the SWIBRID dataset (Figure 17C), indicating that the hypomethylation cannot be considered a feature of J-CH1 insert origins.

Following Lebedin et al., 2022, I analyzed other DNA characteristics that may promote J-CH1 insert donation: R-loops, ERFS and telomeres. R-loops are DNA:RNA hybrids that expose ssDNA of the complementary strand. R-loop formation can cause dsDNA breaks in the exposed ssDNA when not resolved by specialized proteins (*174*). Therefore, I decided to investigate the proximity of the origins of J-CH1 inserts to R-loops using an immunoprecipitation sequencing database called DRIP-seq, derived from a human leukemic cell line (*173*). Moreover, oncogenic stress has been shown to target genomic regions called fragile sites preferentially. This group is divided into common fragile sites and ERFS (*162*). Considering that Lebendin et al., 2022 showed that J-CH1 insert origins were proximal to ERFS, I also decided to analyze ERFS. And last, I decided to analyze the proximity of the SWIBRID J-CH1 insert origin to telomeres. SWIBRID J-CH1 insert origins were closer than the control to the three analyzed datasets: R-loops, ERFS and telomeres (Figure 17D-F). Proximity to these features was in line with the data found in J-CH1 insert origin found in transcripts (*97*).

In conclusion, SWIBRID J-CH1 insert origin was mainly intronic, hypomethylated and close to telomeres, ERFS and R-loops. SWIBRID J-CH1 inserts agree with the published data regarding the proximity but not to the origin signature.

7.4. The DNA repair biomarker – the Prototype

SWIBRID was developed to study J-CH1 inserts; however, the efficiency of J-CH1 insert identification is very low. According to the data, J-CH1 inserts were found in 10 per 1 000 reads (Figure *14*C), leading to the unused portion of 990 reads per 1 000 reads. Luckily, we built up SWIBRID analysis not only recorded data from J-CH1 inserts but also profiled the breakpoints occurring in the switch-joints of the samples. Switch-joints carry the genomic scars of the dsDNA breaks formed during CSR. Since the switch-joints and any other somatic dsDNA break are most likely repaired by cNHEJ, one can consider the switch-joint breakpoint profile to reflect dsDNA break repair in the whole body. Therefore, analysis of the completion of the reads obtained by SWIBRID could help to understand somatic dsDNA break repair in healthy and sick patients. Current tools do not look at the repair event to understand DNA repair but rather at mutations driving DNA repair malfunction. The development of a tool that identifies healthy and DRdef samples based on the switch-joint breakpoint profiles could prognosis diseases related to B cell biology or DNA repair mechanisms with the advantage that we would not miss players of the DNA repair machinery that have not been yet identified. Thus, in this thesis, I decided to create the basis of the **DNA repair biomarker.**

7.4.1. Two-dimensional breakpoint plots reveal specific breakpoint patterns in patient samples

Firstly, to start the switch-joint analysis, the data needed to be visualized. SWIBRID breakpoint analysis of DRdef samples was depicted in a histogram that showed the breakpoints accumulating in the core of the switch regions (x-axis). Breakpoint histograms depicted every class-switching event as two independent events, one in the donor switch region (upstream) and one in the acceptor switch region (downstream). The accumulation of the breaks in SM was specially conserved among WT, $ATM^{-/-}$, $NIPBL^{-/-}$, $BRCA1^{-/-}$ and Ligase $4^{-/-}$, but not for $AID^{-/-}$ (Figure 18A), which suggests that the specificity of the SM breakpoint profile is linked to AID. In addition, the histograms showed that $AID^{-/-}$ and $ATM^{-/-}$ did not contain breakpoints in any SGs (Figure 18A), implying that the successful CSR to SGs requires very early decisions in the molecular cascade of the DNA repair, such as the promotion of dsDNA break, by AID, or sensing of dsDNA breaks, by ATM. Ligase $4^{-/-}$ did not elicit breakpoints in SG2 and SG4 switch regions and had a lower amount of breakpoints than HD in SG3 and SG1 (Figure 18A). No apparent differences were observed between $BRCA1^{-/-}$ and $NIPBL^{-/-}$ compared to HD in the breakpoint histograms (Figure 18A).

The lack of differences between $BRCA1^{-/-}$ and $NIPBL^{-/-}$ compared to HD made me realize that considering every class-switching event as two independent events is insufficient to study DNA repair. The class-switch event results from one repair event in which the DNA repair machinery joins two switch regions. It is crucial to consider the sequence features of the parts that had to be joined to pinpoint specific repairs carried out by certain proteins, e.g., RAD51 would use high stretches of homology to repair the DNA (*175*), only quantifiable when both strands of a break are analyzed together. Consequently, we developed the two-dimensional breakpoint plot (2D-bps plot) that counts every switching event as one considering the exact position of the switch regions involved**.** The 2D-bps plot has a y-axis representing the coordinates of the donor switch regions (upstream) and an x-axis representing the coordinate of the acceptor switch regions (downstream) (Figure 18B). The 2D-bps plot does not consider every nucleotide as a single position but instead uses binsizes to depict and analyze the data. A binsize is determined by grouping nucleotides as units. For instance, if a binsize of 500 is chosen, the sequence would be divided into groups of 500 nucleotides. This affects the analysis, as any nucleotide present in a group can be considered a match in the bin. The 2D-bps plot allowed for the observation of three different types of genomic scars created by CSR. Firstly, direct switch events occurring between SM and another switch region. Secondly, intra-switch deletions are deletions occurring within the same switch region. And third, sequential class-switch events occur between two switch regions that are not SM, likely happening in a second CSR event.

DRdef and HD samples were portrayed in the 2D-bps plots using one read per eff_ncluster and 200 nt bins. The 2D-bps plots allowed us then to differentiate the DRdef from the HD by eye. $ATM^{-/-}$ and Ligase $4^{-/-}$ samples mainly concentrated their breakpoints along SM, SA1 and SA2, and unlike in Figure 18A, the $BRCA1^{-/-}$ sample could be separated from the HD because it lacked intra-switch deletions in SGs. Nonetheless, the 2D-bps plot of $NIPBL^{-/-}$ was similar to HD, suggesting that $NIPBL^{-/-}$ might not disrupt CSR. It has been reported that $NIPBL^{-/-}$ promotes

a-EJ as the DNA repair mechanism during CSR, the second most used DNA repair mechanism in CSR (*176*) (Figure 18C).

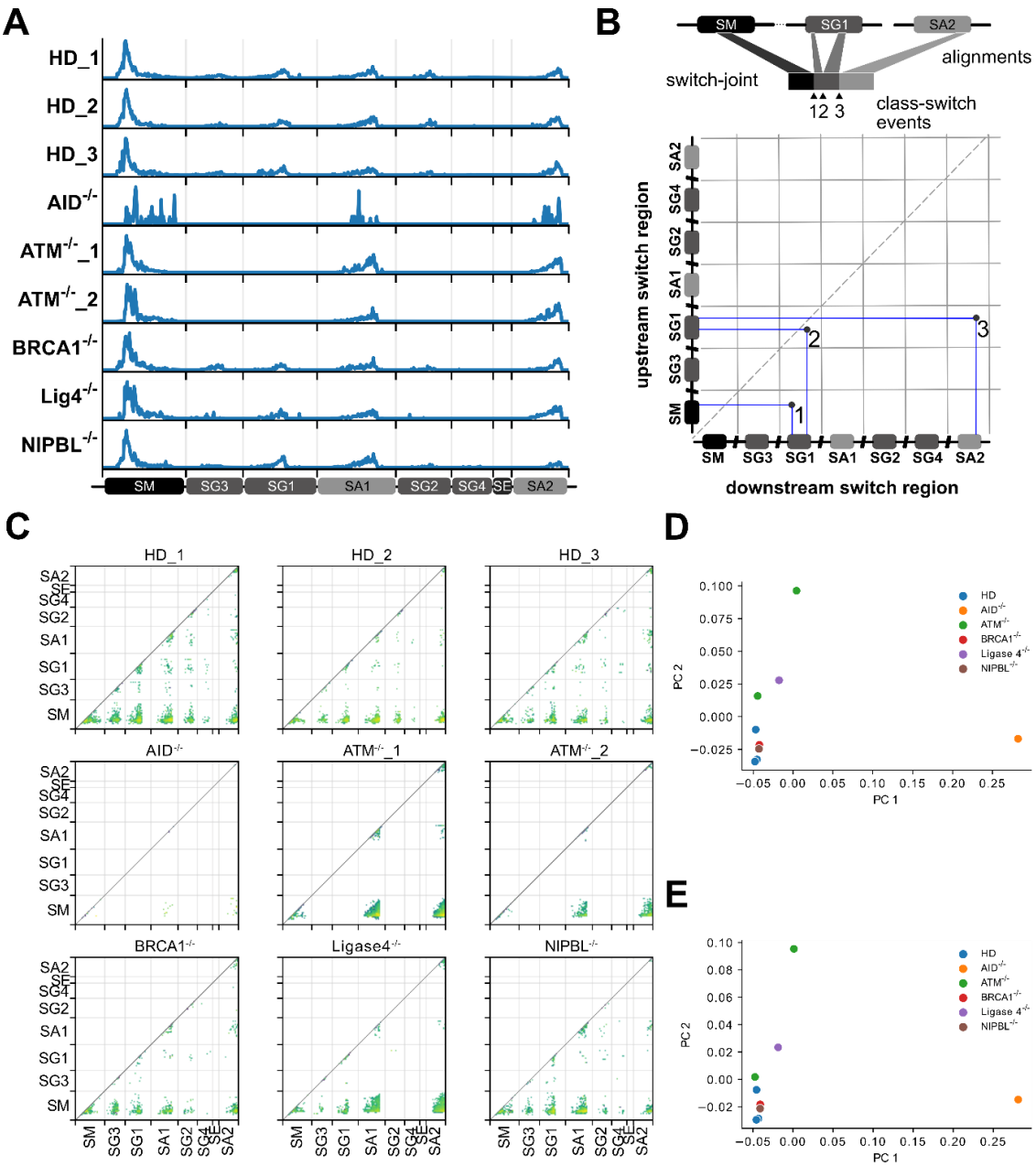


Figure 18. A two-dimensional breakpoint plot of switch-joints allows the differentiation of DNA repair-deficient from healthy samples.

A. Breakpoint histograms from switch-joints of DNA repair-deficient patients and HD in 50 nt bins. Breakpoints are considered every break in the alignment of the switch-joint except primer binding sites. On the left is the sample annotation that corresponds to the histogram. Below are the annotations of the switch regions that align with the breakpoints. Grey boxes indicate the separation between the switch regions. **B.** Shown is the explanation of the two-dimensional breakpoint plot (2D-bps plot). The upper scheme represents a switch-joint containing sequences aligning to SM, SG1 and SA2. Triangles (1,2,3) determine class-switch events. The lower scheme represents a 2D-bps

plot with three dots corresponding to the number in the numbers next to the triangles. Dots are depicted considering the location of the breakpoint in the upstream (y-axis) and downstream switch region (x-axis). **C.** 2D-bps plot of DNA repair-deficient patients (DRdef) and HD using 200 nt bins. These plots show the normalized (per eff_nclusters) density of breakpoints. Color gradients indicate higher or lower breakpoints density. The plots represent the breakpoint in the upstream (y-axis) and downstream switch region (x-axis). **D.** Principal component analysis (PCA) of DRdef and HD using the breakpoint density in the IGH locus and 200-nt bins. Colors indicate different genotypes. PC1 and PC2 are used for depiction in this plot. **E.** PCA of DRdef and HD using the breakpoint density in SM, SA1 and SA2 using 200-nt bins. Colors indicate different genotypes. PC1 and PC2 are used for depiction in this plot. n=1 for all samples.

Principal component analysis (PCA) was chosen to visualize if the event positions in the 2D-plots could spot differences between the DRdef from HD samples. PCA is a linear multi-variable analysis that calculates how distant samples are from each other based on the variables given (PC) and ranks the PCs from the highest to the lowest in terms of their ability to account for the variance between samples (PC1, PC2, PC3 …). Visual separation of samples based on two different PCs would indicate that samples differ based on the variables analyzed. In this case, the PCA considered every bin in the plot a feature, and the number of events occurring in every bin was the value of the feature. These data were collected from the samples and the PCA was performed. PC1 versus PC2, the values that show the highest variability between samples, separated all DRdef from HD except BRCA1$^{-/-}$ and NIPBL$^{-/-}$ (Figure 18D). Potentially, BRCA1$^{-/-}$ and NIBPL$^{-/-}$ samples were clustered close to the HD because they carried breakpoints in SGs, unlike the other DRdef samples. Hence, we analyzed the samples with a PCA using only SM, SA1 and SA2, which did not succeed in separating BRCA1$^{-/-}$ and NIPBL$^{-/-}$ from the other samples (Figure 18E).

In summary, we developed a way to analyze and separate DRdef samples from HD by depicting class-switching events in a two-dimensional plot with axes representing the donor and acceptor switch regions and independent from the switch-joint isotypes of the samples. The separation between DRdef and HD opened a path to create a tool to spot people with DNA repair defects using switch-joints breakpoint profiles. Additional work towards optimizing this tool will eventually lead to the development of the **DNA repair biomarker**.

7.4.2. Translation of the 2D-bps plots to mouse samples with an advantage to use more DNA repair deficient samples

Human B cells contain a high diversity in their switch-joints since humans are constantly exposed to pathogens pushing B cell diversification to generate highly variable antibodies. The variability is a sum of the different isotypes found in the switch-joint library and the randomness of the breakpoints throughout the switch regions. Since the human IGH contains 8 isotypes and is ~200 Kb long (Figure *6*B), the variability among switch-joints can be immensely high. Thus, defining a healthy DNA break repair profile is challenging, and a large cohort should be used

to analyze breakpoint profiles in switch-joints. Optimally, we would use DRdef human samples to establish the DNA repair biomarker in humans, but unfortunately, DRdef humans are scarce and difficult to obtain. Hence, we decided to determine the DNA repair biomarker in mice because we could access CH12 knockout cells (Table 2) and generate as many replicates as necessary. Considering that it is a cell line, we decided to generate several replicates of the same CH12 knockout genotype and analyze them with SWIBRID to estimate whether the switch-joints breakpoint profiles could separate the different knockout genotypes.

Previous analysis on human 2D-bps plots showed that PCA only using SM, SA1, and SA2 did not influence the separation between samples (Figure 18E). Therefore, performing PCA on 2D-bps plots of CH12 samples whose breakpoints would be mainly focused between SM and SA would not influence the separation of the different CH12 knockout genotypes.

In contrast to polyclonal human samples, activation of CH12 generated a lower number of eff_nclusters. Firstly, because they mostly switch to SA, and secondly, they are a B cell line with a common ancestor and might class-switch using a pattern. I also decided to run WT Peyer´s patches (PP) samples to compare the diversity found in CH12 to a mouse sample. Due to the lower number of eff_nclusters, the visualization of the 2D-bps plots became less apparent. Hence, we decided to use a binsize 200 for visualization, while the analysis of the 2D-bps plot was done using a binsize 50 to comply with the analysis done in human samples (Figure *19*A). Visualization of CH12 and PP using 2D-bps plots showed a clear difference in IGH locus usage. Keeping in mind that the switch-joints analyzed in mouse samples were SM-SG3, SM-SG2b, SM-SG2c and SM-SA, CH12 mainly showed breakpoints focused on the SM-SA area, while PP showed a higher diversity of breakpoints along the whole IGH locus with a higher frequency in SM-SA. Unfortunately, PP was not highly exploited in this study due to the difficulty of obtaining mouse knockouts at this point (Figure *19*B). The comparison of CH12 2D-bps plots shows that WT and BRCA1$^{-/-}$ have very similar breakpoint profiles in SM-SA and SM-SG2b, although CH12 WT has higher breakpoints in the latter. CH12 Ligase 4$^{-/-}$ also had a majority of breakpoints in the SM-SA zone but was also scarce in the SM-SG2b zone. CH12 53BP1$^{-/-}$ and Rif1$^{-/-}$ has almost no breakpoints and the few ones they had were focused on the SM-SA zone (Figure *19*B).

2D-bps plot PCA of CH12 samples standardized by eff_ncluster and a binsize of 50 did not reveal a separation between the different CH12 knockout genotypes. In an attempt to study the core problem of the lack of separation between the knockouts and the wild-type samples, we visualized the genotypes (WT, Rif1$^{-/-}$, BRCA1$^{-/-}$, Ligase 4$^{-/-}$ and 53BP1$^{-/-}$), the MinION run date (batch 1, batch 2, batch 3), the eff_nclusters and the number of reads in each sample (nreads) in a PCA plot (PC1 versus PC2) (Figure *19*C). Comparison between these features showed that the date of the MinION run had the highest impact since different batches appeared clustered independently of the CH12 genotype (Figure *19*C-Batch).

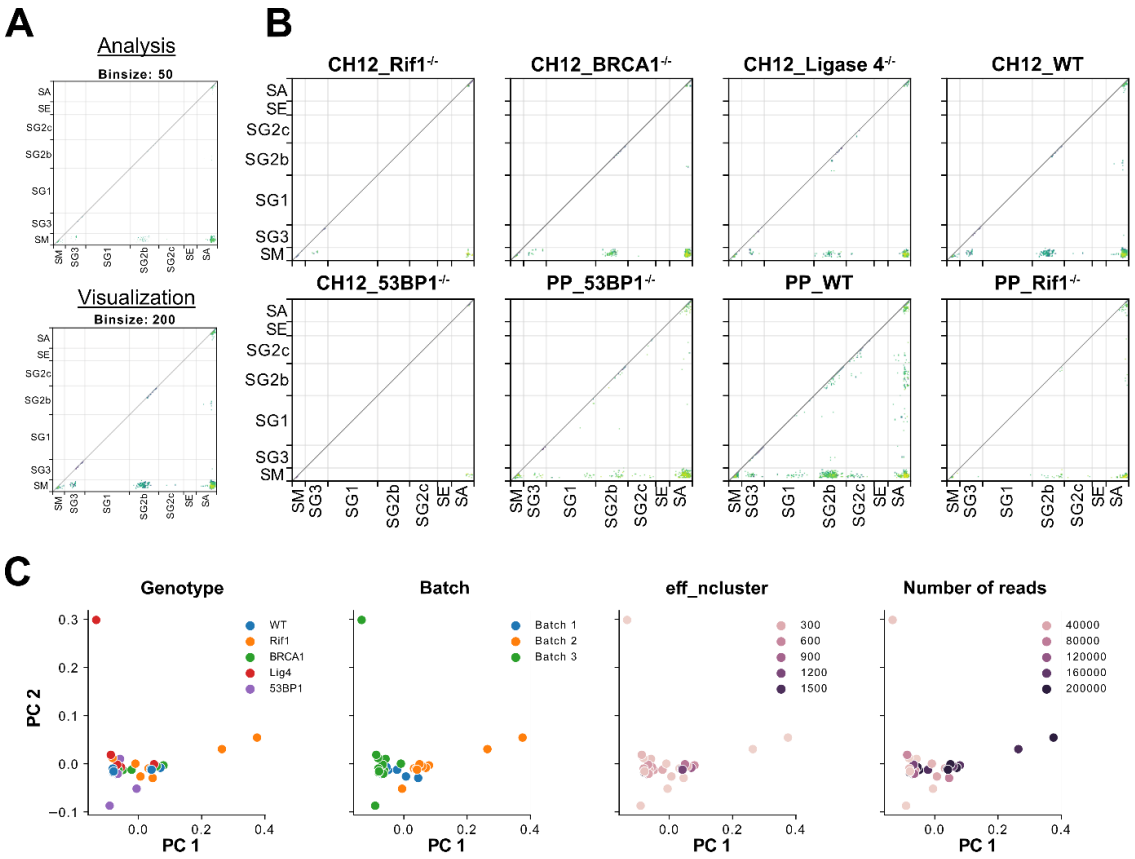


Figure 19. Translation of two-dimensional breakpoint plots in murine samples allows us to identify the limitations of SWIBRID.

A. Explanation of the binsize used for analysis and visualization. Binsize=50 was used to perform the PCA analysis, and binsize=200 was used to visualize the two-dimensional breakpoint plots (2D-bps plots). **B.** 2D-bps plots of standardized murine samples (by eff_nclusters) using a binsize of 200. Rif1, 53BP1 and Ligase 4 are proteins from the cNHEJ DNA repair mechanism. BRCA1 is a protein from the homologous recombination DNA repair mechanism. Rif1=Replication timing regulatory factor 1, 53BP1=p53 binding protein 1, BRCA1=Breast cancer gene 1, WT=Wild-type, CH12=B cell lymphoma cell line. PP=Peyer's patches. In A and B, dots are depicted considering the location of the breakpoint in the upstream (y-axis) and downstream switch region (x-axis). **C.** Principal component analysis (PCA) of CH12 samples carrying different knockouts (PC1 versus PC2). Knockouts are indicated in the graph in the "genotype" plot. Legends of the plots are located in the right of the PCA plots. The title of the plot indicates what the color coding of the legend shows. Technical replicates in CH12 WT, n=6; CH12 Rif1$^{-/-}$, n=7; CH12 BRCA1$^{-/-}$, n=6; CH12 Ligase 4$^{-/-}$, n=6; CH12 53BP1$^{-/-}$, n=4.

In conclusion, analysis of the 2D-bps plot using CH12 knockouts allowed us to use a higher number of replicates per genotype and revealed that the date of the MinION run comprised the most significant limitation of SWIBRID, which we

would not be able to spot by the single human DRdef samples. The development of a DNA repair biomarker would need a deconvolution of the 2D-bps plots into single characteristics.

7.4.3. Exploiting the 2D-bps plots in human samples to develop the DNA repair biomarker in humans

Analysis of the switch-joint breakpoints using only the bins as features did not differentiate the healthy from the disease genotypes in humans or mice. Therefore, we decided to deeply characterize the genomic scars considering crucial characteristics of the DNA repair and CSR events.

The 2D-bps plot depiction can classify the breakpoints at different levels: single-breakpoints, breakpoints occurring between SM and any other switch region, multiple-breakpoints, breakpoints occurring between two switch regions that are not SM (as a result of sequential switching) and within-breakpoints, meaning breakpoints that occur intra-switch region deletions (Figure 20A). Classification of breakpoints is potentially crucial to separate DRdef from HD because DNA breaks in DRdef patients are more likely to be left unrepaired, which threatens the life of the B cells (*177*). Thus, DRdef B cells would likely not carry as many breaks as HD because the ones with breaks would probably die.

Analysis of breakpoint classes in DRdef and 20 HD (same sample material, PBMC) showed a tendency of a higher occurrence of single-breakpoints in DRdef samples but higher within-breakpoints and multiple-breakpoints in HD (Figure 20B). Interestingly, similar tendencies in the breakpoint class were found among HD of different ethnicities (European and African) and between different cell number samples from HD4 (Figure 20B-C), which speaks for the conservation of these features along HD and within the same donor. More replicates will be necessary to build significance in DRdef samples versus HD (DRdef n=6).

With the purpose of finding specific features to separate healthy from DRdef samples, I decided to record the number of inversions occurring in a sample per eff_ncluster. Inversions are CSR events in which the joining between a (-)-sense switch region and a (+)-sense switch region occurs (*178*, *179*). Inversion frequency was found to occur significantly more frequently in HD than in $ATM^{-/-}$ and Ligase $4^{-/-}$ samples, while the frequency was similar comparing European and African donors (Figure 20D). These results agree with the fact that 53BP1, a core protein in the DNA repair machinery during CSR working together with ATM and Ligase 4, has been found to regulate inversional joins tightly (*78*).

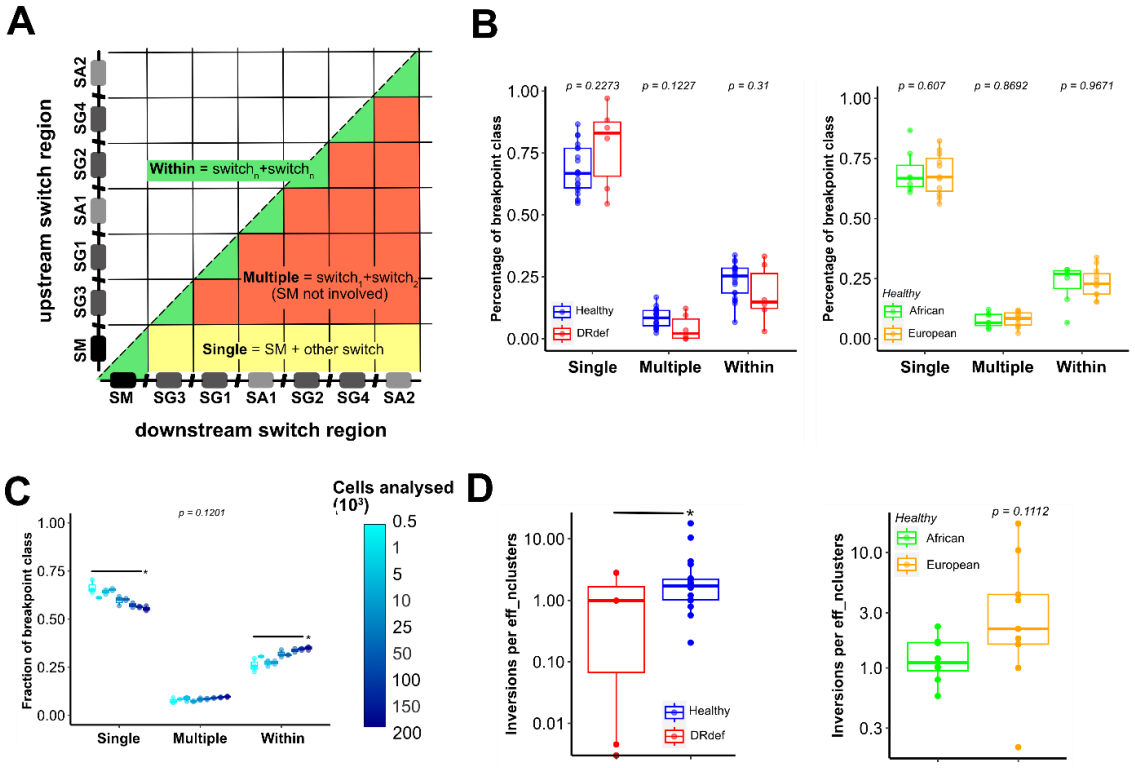


Figure 20. DNA repair protein deficiencies show a different inversion frequency but not in breakpoint class distribution.

A. Scheme of breakpoint classification using the 2D-bps plot. Areas described as single-, multiple-, and within-breakpoints are depicted in the 2D-bps plot in yellow, orange and green, respectively. X-axis represents the breakpoints in downstream switch regions and the y-axis represents the breakpoints in upstream switch regions. **B.** Difference of multiple-, single- and within-breakpoint frequency between DNA repair protein deficient patients (DRdef) and healthy donors of PBMC material origin (left) and donors with an African and European background (right). **C.** Breakpoint class distribution along a different sampling of the HD4. Three technical replicates were performed for each cell number analyzed. Different shades of blue represent different amounts of cells. **D.** Inversions per eff_nclusters: DRdef versus HD of PBMC material origin (left) and Africans and Europeans (right). Two technical replicates were performed for each sample: healthy African donors, n=5; healthy European donors, n=5; Healthy, n=20 and DRdef, n=6. HD=Healthy donors. T-test was performed after the normal distribution of the data was confirmed through the Shapiro test. *=p-value < 0.05.

Furthermore**,** DNA repair protein deficiencies trigger the replacement of preferred DNA repair mechanisms (*176*, *180*). The preferred DNA repair mechanisms during CSR are c-NHEJ and a-EJ, which often use <20 nt homologies. Homologous repair (HR) or single-strand annealing (SSA), which work exclusively using >20 nt homologies (*69*), can take over the repair of dsDNA breaks when cNHEJ or aEJ cannot be used. Since such a different homology

range is found in homology and non-homology DNA repair mechanisms, we analyzed the average sequence homology used for switch-joint repair.

The homology was calculated by measuring the times the repaired strands in a break shared the same nt at the point of the break. Instead of comparing the strands nt by nt, the analysis was done in bins of 4 nt; thus, we called it homology 4-mer. In this analysis, raw sequences were not directly compared since MinION errors could lead to false results. As an alternative and trusting SWIBRID's alignment of the libraries to the IGH, the homology 4-mer was calculated using the coordinates of the reads, specifically, the 2D-bps plot. The homology found in 4 bins was depicted in a plot with the same configuration (x- and y-axis) as the 2D-bps plot (Figure *21*A). The homology plot showed the most homologous sequences in red and the least in white. The homology 4-mer in samples comprises the fraction of 50 bin breaks that occurred in 4-bin homologous zones. Furthermore, we differentiated between the homology 4-mer occurring in two strands following the same sense, homology (-) 4-mer (Figure *21*A-lower triangle), and two strands of opposite sense, homology (+/-) 4-mer (Figure *21*A-upper triangle). The differentiation was necessary to consider the occurrence of inversions during CSR. Results show that switch-joints in DRdef used significantly higher homology (-) 4-mer than HD, but the opposite occurred for homology (+/-) 4-mer (Figure *21*B). Specifically, ATM$^{-/-}$, AID$^{-/-}$ and Ligase 4$^{-/-}$ had the highest differences from HD. The data agree that ATM$^{-/-}$ and Ligase 4$^{-/-}$ do not experience many inversions (Figure 20D). AID-/- data in inversions is missing due to the lack of clusters.

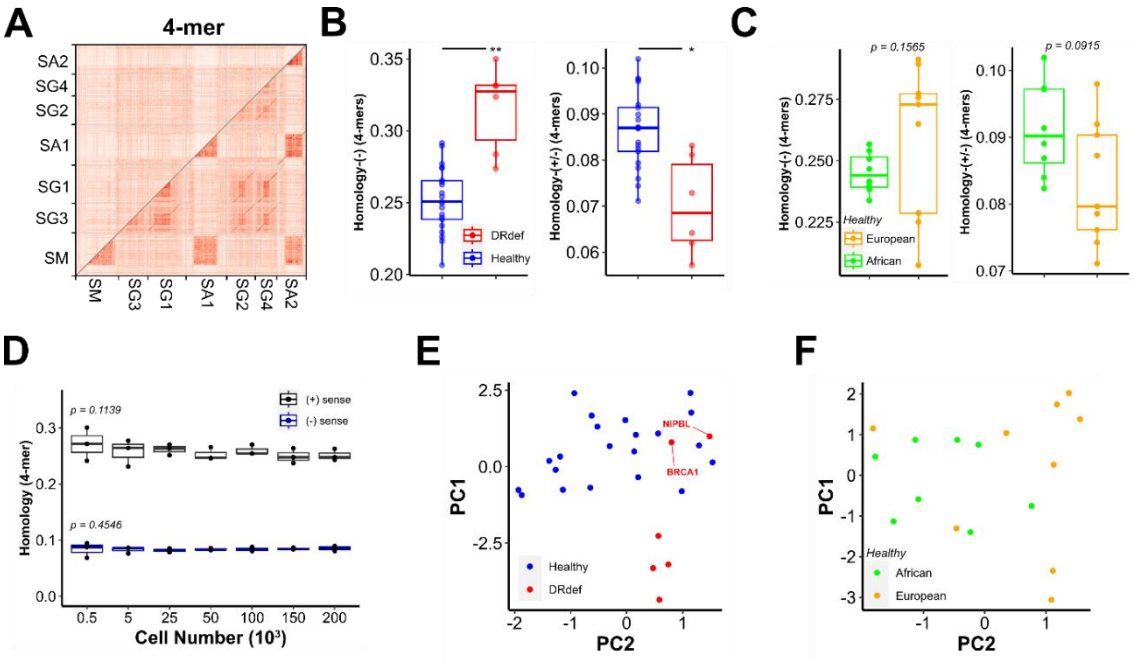


Figure 21. DNA repair deficiencies show different usage of 4-nt homologies when switching.

A. Scheme of the 4-mer homology density between switch regions along the IGH. The 4-mer homology was calculated using 4 nt bins. For every homology 4-mer bin, a red square was depicted. The lower triangle depicts the homology occurring between (-)-sense sequences in the IGH, while the upper triangle depicts the homology occurring between (+)- and (-)-sense sequences in the IGH. **B-C.** Average sequence homology used in switching between DNA repair protein deficient patients (DRdef) and HD (**B**) and healthy donors with African and European backgrounds (**C**). Homology between the (-) sense on the left and between the (+/-) sense on the right. **D.** Percentage of homology used in switching among different cell number samples of HD4. Three replicates per cell amount were analyzed. T-test was performed after the normal distribution of the data was confirmed by the Shapiro test. **=p-value < 0.01. **E-F.** Principal component analysis (PCA) DRdef versus HD (**E**) and healthy donors with African versus European background (**F**). Variables used for PCA: homology (+/-) sense, homology (-) sense, log10(inversions per eff_nclusters) and single-, within- and multiple-breakpoints. Variables were scaled before the PCA. Analysis performed under seed 147. Healthy, n=20; DRdef, n=6; African donors, n=5; European donors, n=5. HD=healthy donor.

We could not consistently separate DRdef from HD using the previous features separately. Therefore, we decided to perform the PCA using the scaled breakpoint classes, inversion frequency, homology (-) 4-mer and homology (+/-) 4-mer. The PCA separated all DRdef from HD except for NIPBL$^{-/-}$ and BRCA1$^{-/-}$ (Figure *21*E). This result was similar to the PCA on the 2D-bps plot when using all isotypes (Figure 18D). NIPBL$^{-/-}$ and BRCA1$^{-/-}$ cluster with HD, most likely because none of the features alone showed any difference between these two deficiencies and HD. As a control, PCA in European and African samples did not show such a big separation. Nevertheless, some samples with European backgrounds were clustering on the edge of the plot (Figure *21*F).

In summary, PCA of specific features of the 2D-bps plots can differentiate ATM$^{-/-}$, AID$^{-/-}$ and Ligase 4$^{-/-}$ samples from HD. NIPBL$^{-/-}$ and BRCA1$^{-/-}$ had similar values to HD in the features analyzed and, therefore, did not separate in the PCA. Including further 2D-bps plot features would help the separation of NIPBL$^{-/-}$ and BRCA1$^{-/-}$. The potential of this analysis would be enhanced by adding more replicates from the same DRdef type, leading to a better separation of DRdef samples from HD samples.

7.4.4. Exploiting the 2D-bps plots in mouse samples to develop the DNA repair biomarker in mice

The main hurdle in optimizing SWIBRID is the low availability of human samples with DNA repair defects. Since much DNA repair research is done in cell lines, we decided to optimize SWIBRID using CH12 cells. Again, we made use of the CH12 knockout cells to investigate the new deep analysis considering features beyond the breakpoints.

CH12 WT, 53BP1$^{-/-}$, BRCA1$^{-/-}$, Ligase 4$^{-/-}$ and Rif1$^{-/-}$ cells were cultured for two days in two rounds, generating 10, 5, 6, 7 and 8 replicates, respectively. SM-SA, SM-SG3 and SM-SG2bc switch-joints were amplified, sequenced by MinION and ran in SWIBRID. Following the human analysis, we calculated the classification of the breakpoints, inversion frequency and homology (-) 4-mer and (+/-) 4-mer. Comparison of breakpoint classification showed a general lack of multiple-breakpoints, or sequential switching, and a significant difference in single- and within-breakpoint frequency between WT and BRCA1$^{-/-}$ CH12 samples (Figure *22*A). Ligase 4$^{-/-}$ had a significantly higher inversion frequency than any other CH12 sample (mean inversion per eff_nclusters: Ligase 4=25.8, BRCA1=0.92, Rif1=2.81, 53BP1=2.06, WT=1.3) (Figure *22*B). Homology density plots were also done as previously described (Figure *22*C); however, unlike in human samples, CH12 samples did not show a different average sequence homology (-) 4-mer or (+/-) 4-mer (Figure *22*D).

Due to the lack of apparent differences among the knockout samples, a new feature was added to the analysis: the spread in the switch regions. The spread represents the average width of the switch regions covered by breakpoint samples. The aim was to add only the spread switch regions that would add differences. Spread in SM and SA was very similar among all samples; however, spread in SG3 showed significant differences between WT and Ligase 4$^{-/-}$ samples and SG2b showed different tendencies among all genotypes (Figure *22*E). Considering the results above, PCA was done using the variables that showed a higher difference (significantly or not) among the genotypes: single-breakpoints, within-breakpoints, inversion per eff_nclusters, spread in SG3 and SG2b. Depiction of PC1 and PC3 clustered the different mutants together. Ligase 4$^{-/-}$ samples clustered separated by at a certain distance from the other samples, whereas 53BP1$^{-/-}$ and Rif1$^{-/-}$ mainly clustered together. Clustering of 53BP1$^{-/-}$ and Rif1$^{-/-}$ together might signify that similar switch-joints result from a knockout of proteins that promote c-NHEJ only when working together (Figure *22*F).

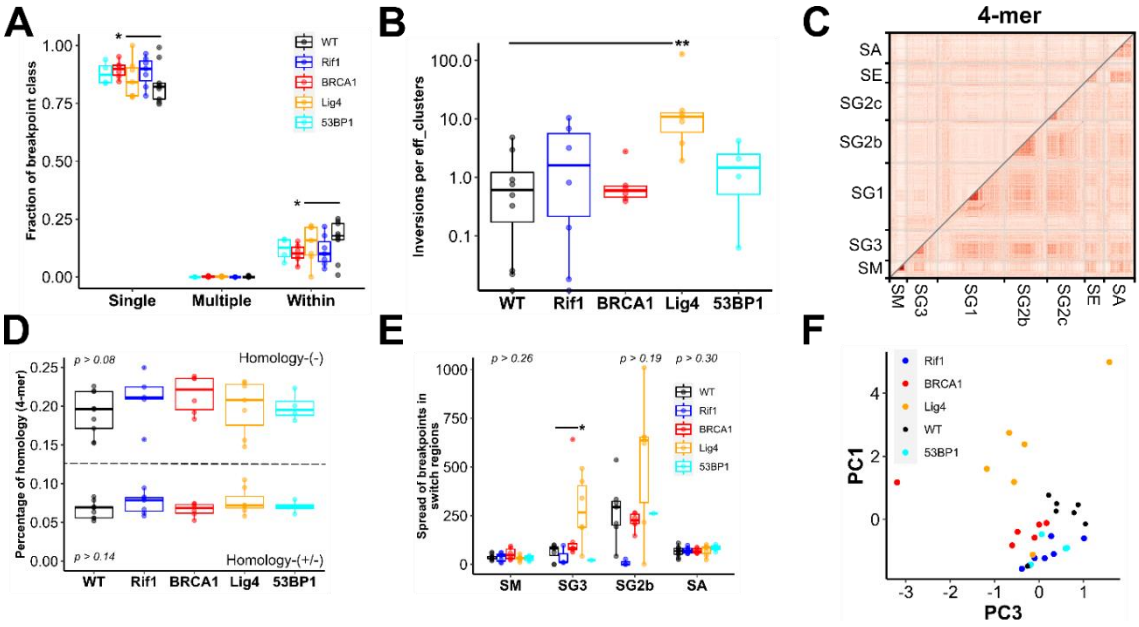


Figure 22. Knockout of Ligase 4 and BRCA1 change the switch-joint profiling in CH12.

A. Difference of multiple-, single-, and within-breakpoint frequency between DNA repair knockouts and wild-type CH12 samples. **B.** Inversions per eff_nclusters among DNA repair protein knockouts in CH12. **C.** Scheme of the 4-mer homology density between switch regions along the IGH. For every homology 4-mer nucleotide, a red square was depicted. The lower triangle depicts the homology occurring between (-)-sense sequences in the IGH, while the upper triangle depicts the homology occurring between (+)- and (-)-sense sequences in the IGH. 4-mer means the homology analysis has been done between bins of 4 nucleotides. **D.** Difference of average sequence homology used among different DNA repair protein knockouts in CH12 between (-)-sense sequences (homology-(-), upper) and between (+)- and (-)-sense sequences (homology-(+/-), lower). T-test was performed after the normal distribution of the data was confirmed by the Shapiro test. *=p-value < 0.05, **=< 0.01. **E.** Spread of breakpoints in SM, SG3, SG2b and SA among DNA repair protein knockouts in CH12. Spread indicates the average width of the switch regions covered by a sample. Normality was not found in the SG3 data distribution and the wilcox.test was performed, *=p-value < 0.05. **F.** PCA of DNA repair protein knockouts and wild-type CH12 samples. PCA was performed using scaled log10 inversions per eff_nclusters, single- and within-breakpoints, and spread in SG3 and SG2b. Technical replicates for CH12 WT, n=10; Rif1$^{-/-}$, n=8; 53BP1$^{-/-}$, n=5; BRCA1$^{-/-}$, n=6 and Ligase 4$^{-/-}$, n=7.

In summary, using features of the 2D-bps plot in CH12 helps to cluster the different CH12 genotypes separately using a PCA. The success of separately clustering CH12 knockouts forms the basis of the DNA repair biomarker, which eventually could be used in human samples to predict disease.

8. DISCUSSION

8.1. B cell switch-joints constitute a biomarker or DNA repair malfunction

The present work introduces SWIBRID, a novel tool to identify DNA repair malfunction by characterizing the breakpoint profiles occurring in switch-joints of B cells upon diversification to describe the switch-joints of a sample comprehensively.

8.1.1. SWIBRID analysis finds mutations more frequently than published methods, and it could be used for the prediction of cancer

In SWIBRID, I studied the mutational break repair of CSR, processed by cNHEJ and aEJ, and could not record the error-free break repair processed by HR. By definition, the latter includes the copy of entire DNA sequences in trans and is error-free. When HR is active during CSR, it does not result in a genomic scar and, therefore, cannot be detected by SWIBRID (*181*). On the other hand, the mutational break repair results in mutations, in the case of CSR, a deletion in the form of a switch-joint in the IGH locus, allowing the amplification of switch-joints (*69*).

Unrepaired DNA breaks trigger apoptosis, which suggests that the cell intrinsically seeks to repair a DNA break rather than leaving it broken (*182*). Indeed, if a DNA repair pathway misses a protein, it will compensate for it using other proteins. This compensation mechanism can potentially repair a DNA break into a threatening outcome. For instance, the deficiency of an HR protein (e.g., BRCA1) would lead to a higher rate of mutations since the DNA break would be repaired more often by a mutational DNA repair pathway (*180*, *183*). In this case, SWIBRID could identify HR impairment since, theoretically, it would record more mutations or aberrations than in samples with healthy HR. Thus, SWIBRID has the potential to identify DNA repair malfunction occurring in any DNA repair pathway, using the switch-joints as a sink of somatic DNA breaks. In this thesis, I was specifically able to separate WT, BRCA1$^{-/-}$, Ligase 4$^{-/-}$, 53BP1$^{-/-}$ and Rif1$^{-/-}$ CH12 using PCA and including the features: direct switch-joint and intra-switch deletion breakpoint (single and within, respectively), inversion per unique B cell clone (eff_nclusters) and the scope of breakpoint occurrence (spread) in SG3 and SG2b. Notably, mutations in the gene of BRCA1, Ligase 4 and 53BP1 in humans are linked to the development of cancer (*184*), which indicates that SWIBRID has the potential to identify different DNA repair malfunctions and, thus, cancer predisposition in human samples.

State-of-the-art in cancer prognosis has commonly screened mutations in DNA repair proteins which can lead to wrong assumptions since gene mutations are not always linked to protein malfunction (*184*). DNA repair efficiency has been measured by other people using direct methods applied by *in vitro* assays that give a semiquantitative measure of genomic DNA breaks in cells exposed to DNA-damaging reagents (comet assay) (*185*). However, it is a quantitative measure of DNA break in the cells and does not necessarily reflect the DNA repair power in a cell. Research on mutations that can initiate or maintain cancer,

also called cancer-driving mutations, has been performed unbiasedly and comprehensively by people like Sherman et al., 2022, who used genome-wide DNA break repair analysis to predict potential cancer-driven mutations in coding and non-coding sequences. Specifically, Sherman et al., 2022 analyzed public whole-genome, whole-exome and targeting sequencing datasets of 37 cancer types using a deep learning approach called Dig. Dig generated a neural network that could identify positively selected cancer-driving mutations of arbitrary cancer cohorts (*141*). Dig allowed the identification of new cancer-driving mutations linked to DNA repair function. For instance, E74 Like ETS Transcription Factor 3 (ELF3), a transcription factor involved in HR (*186*), is an oncogene whose malignancy has been previously linked to its gene variants (*187*). Nonetheless, Dig novelly found mutations in the ELF3 promoter that were suggested to enable its methylation and expression in two cancer cohorts (*141*). Cancer-driving mutations are crucial to identify, screen for, and asses for people's predisposition to cancer. However, screening all cancer-driving mutations found in Sherman et al., 2022 would involve the analysis of at least a whole-exome sequencing, which is not an accessible screening to perform. Therefore, the approach of SWIBRID to analyze DNA repair function and, indirectly, cancer predisposition, using only the breakpoint profile of switch-joints located in one gene, is feasible to envisage in the clinics as a predisposition analysis since it works amplifying a focused area and can be used using a focused computational analysis.

DNA repair research commonly uses B cells due to the sink for DNA-break events in the IGH as a result of CSR that frequently become somatic DNA lesions. Following these lines, we analyzed the switch-joints from B cells, assuming that switch-joints mirror the somatic DNA repair in the body. Studies have considered that a successful CSR in B cell proves that cNHEJ or aEJ works. For instance, the role of 53BP1 (*77*) or Ligase 4 (*188*) for end-joining DNA repair in splenic murine B cells activated *in vitro* followed by the analysis of switch-joints or in CH12 activated *in vitro* followed by the analysis of IgA switching, respectively. Recently, it has been shown that the mutational signatures occurring during antibody diversification are the same types that trigger B cell lymphoma, which links somatic DNA repair in B cells to cancer. Specifically, Machado et al., 2022 analyzed the mutational landscape of B cells and T cells in different subpopulations and compared the mutational burden found in whole-genome sequencing in four HD to their hematopoietic stem and progenitor cells (HSPCs). Lymphoid malignancies, such as chronic lymphocytic leukemia and malignant B cell lymphomas, is characterized by single base substitution signature 9 (SBS9), which are T>G transversions at ApTpN and TpTpN trinucleotides and have been proposed to be linked to the activity of the error-prone Pol η (*189*). Memory B cells of lymphoma patients show a direct link between the SHM rate and the number of SBS9 mutations found in the genome (1 SHM: 18 SBS9) (*190*). This fact links the mutations occurring during B cell diversification to those promoting cancer located around the whole genome. More importantly, it has been found that neurons and oligodendrocytes accumulate mutations with age and follow the same mutational signatures as the HD HSPCs cells that Machado et al., 2022 observed (*191*). Neurons and HSPCs, two completely different cell types, similarly mutate their genome, meaning that they similarly repair their DNA

breaks. Therefore, we can suggest that the switch-joints in B cells could resemble or mirror the DNA repair throughout the body.

Sherman et al., 2022 and Machado et al., 2022 show unbiased studies that analyzed either the predisposition to cancer or the mutational burden using whole-genome sequencing because they did not focus on specific genomic regions. However, the disadvantage of using such a methodology is the low likelihood of finding a mutation. Sherman et al., 2022 used publicly available sequencing data from cancer cohorts. They looked for mutations using 10 Kb-bins (groups of 10 kb nucleotides), allowing them to find a mutation per 1.3 bins analyzed (~ 1 mutation per 13 Kb, in 75% of the bins). In contrast, Machado et al., 2022 sequenced the whole genome using 3000 cells per subtype (T cell, B cell, HSPCs). The samples consisted of an expansion of individual clones belonging to different cell subtypes. They looked for mutations also using 10 Kb bins, binning the genome in 279 094 bins and finding only 91 343 bins carrying mutations; therefore, they only found mutations in 32% of the bins (*141*, *190*). In contrast to these studies, the present work takes advantage of the fact that numerous DNA breaks occur in close proximity leaving the whole genome analysis unnecessary and finding at least one mutation corresponding to the CSR event in every read. Sequencing and analyzing reads of a maximum of 5 kb instead of whole chromosomes makes the pipeline require less computational power as well as a more detailed characterization of a region in terms of somatic breaks.

In conclusion, SWIBRID presents a tool to identify cancer predisposition by analyzing the efficiency of DNA repair in switch-joints. In contrast to current efforts to predispose cancer focus that search for cancer-driven mutations or mutational signatures using whole-genome sequencing, SWIBRID screens the efficiency of somatic DNA repair. Also, whole-genome sequencing-based approaches can identify relevant information in as much as 75% of the analyzed data. In turn, SWIBRID identifies at least one genomic scar to analyze in every read, translating into a highly efficient tool to identify the risk of cancer development.

8.1.2. DNA repair analysis by SWIBRID has the potential to identify neurological and immune-related diseases

Analysis of DNA repair function provides information about a patient's chance of repairing a break error-free compared to a healthy sample. Besides cancer, DNA repair malfunction is associated with the development of neurological diseases. For instance, mutations in ATM can lead to the development of ataxia–telangiectasia, a neurodegenerative disease that causes ataxia early in life (*192*). DNA repair analysis also focuses on screening the function of DNA repair proteins, although many DNA repair factors remain unknown. Unlike present SNP screening methods (*193*), the DNA repair biomarker could potentially predict neurodevelopmental diseases associated with DNA repair deficiencies early in life.

Moreover, T and B cell diversification relies on the imprecise gene recombination of their TCR and BCR. These recombinations require mutational DNA repair

mechanisms such as cNHEJ to repair the joints (*194*). Therefore, it has been shown that impaired DNA repair proteins in humans can result in primary immunodeficiencies. In fact, Ligase 4 deficiency is the cause of Ligase IV Syndrome, a T-B-Natural Killer (NK)+ severe combined immunodeficiency (*195*). Most immunodeficiencies are challenging to diagnose since their signs can be easily confused with other diseases like sinusitis. Currently, diagnosing B cell-related primary immunodeficiencies includes the analysis of clinical signs, such as recurrent infections or antibody classes in the blood. Genomic screening is not often used as diagnostics, given that nuclear factor-kappa B subunit 1 (*NFKB1)*, the most common defect causing common variable immunodeficiency (CVID), is only found in 4% of the patients (*196*). Thaventhiran et al., 2020 performed a genome-wide association study (GWAS) on 974 patients carrying sporadic or familiar CVID with a large control dataset in pursuit of novel common genomic defects. Unfortunately, they did not find a CVID genomic defect more common than nuclear factor kappa B subunit 1 (*NFKB1)*, but they found new ones that the Exome Aggregation Consortium did not cover before, such as deletion on Actin Related Protein 2/3 Complex Subunit 1B (*ARPC1B)* that impaired the assembly of actin in immune cells (*197*). These studies suggest that the discovery of new genetic variants could facilitate accurate diagnosis of primary B cell immunodeficiencies. However, mutations identified using GWAS are not necessarily linked to a phenotype, and the consequence of a gene mutation should be further investigated at the protein level. On the other hand, SWIBRID profiles repair outcomes in DNA repair in genomic scars rather than examining the functionality of specific molecular players, identifying differences in DNA break repair between a healthy and immunodeficient patient and linking it to a phenotype.

SWIBRID could help diagnose immunodeficiencies such as IgA deficiency in children, the most prevalent immunodeficiency in the Western world (*198*) since blood IgA level in children younger than 4 years old is scarce and the deficiency complicates the diagnosis at this age. IgA deficiency can undergo unnoticed and not cause any consequence, but in some untreated cases, it can cause severe complications. Clinical cases show a 4 and 2 years old with an undiagnosed IgA deficiency which caused nephropathy (*199*), meaning that IgA deficiency also affects children that, in principle, should have very low IgA in their blood. Since IgA switching is not impaired in children, considering it can be found in saliva at higher concentrations than in blood (*200*), SWIBRID could determine the switching percentage from blood and help diagnose IgA deficiency in children younger than 4 years old.

Therefore, the phenotypic analysis of switch-joints may be preferable over genotyping for diagnosis. It would be of utmost importance to detect them early in life, as some primary immunodeficiencies are not expressed symptomatically until adulthood, threatening the life of the patients with the development of disease complications such as B cell lymphoma (*201*). SWIBRID analyzes the switch-joints of B cells, which also allows the identification of specific immunodeficiencies based on their isotype switching percentage, such as hyper IgE (*202*) or IgA deficiency (*203*). Thus, it can observe impaired B cell

diversification or switching – two leading causes of B cell primary immunodeficiencies (according to the European Society for Immunodeficiencies Registry) – likely allowing their early-life diagnosis.

In conclusion, the DNA repair biomarker has the potential to identify DNA repair malfunction using the switch-joints and predict the development of neurodevelopmental diseases and immunodeficiencies that currently are challenging to diagnose.

8.2. SWIBRID comprises a new methodology to characterize switch-joints in humans and mice using novel features

The study of CSR by switch-joint analysis started in 1991 when SM-SE switch-joints were PCR amplified to disprove that IgE expression occurred as a result of RNA splicing (*204*). Recently, switch-joints were used to discover that the combination of the shieldin complex subunit 1 (SHLD1) and XRCC4-like factor 1 (Xlf1) was crucial for the inhibition of resection (*205*) or to confirm that staggered dsDNA breaks are often repaired using resection and MH (*206*). Here I compare the switch-joint analysis of SWIBRID with published switch-joint data in humans and mice.

Switch-joint analysis has been commonly standardized by the total number of reads or not normalized at all (*178*, *205*). However, standardizing by the number of reads is not optimal because clones in a polyclonal library are represented by one or more reads. For instance, all CH12 knockouts, especially Rif1$^{-/-}$, presented less diversity than WT CH12. Standardizing our analysis using reads gives an inversion frequency in Rif1$^{-/-}$ CH12 of 0.0044, and doing it with eff_nclusters is 2.81. DNA repair protein inhibition analysis is expected to result in a lower B cell diversity. Therefore, standardizing the switch-joint analysis using B cell clones is essential to avoid overestimating events in samples with low diversity (*207*, *208*). By developing SWIBRID, we have created a tool that not only studies the DNA repair of breaks during CSR and quantifies the number of individual B cell clones in a sample, for the first time enabling the standardized quantification of DNA repair in switch-joints. Thus, the standardization of the analysis using clones is primordial.

To underline the relevance of quantifying B cell clones for switch-joint analysis, we set out to reanalyze a published dataset using SWIBRID. Vincendeau et al., 2022 sequenced SM-SG1 joints from murine splenic B cells by PacBio and analyzed the resection occurring in amplicons standardizing their data using the number of reads. When applying SWIBRID to the dataset, we observed fewer and unequal unique B cell clusters depending on the sample. For instance, the WT sample contained 1 276 reads, while SWIBRID identified only 174 unique B cell clones. Thus, SWIBRID has a significant advantage of standardizing the data in a high-throughput manner by the number of individual B cell clones, which is especially crucial when analyzing samples with different B cell diversity.

Homology is a feature that is often measured in switch-joint analysis. It is described as the similarity between two sequences and a feature that can help

determine the DNA repair mechanism involved in DNA break repair (*69*). On the one hand, our human results show that DNA repair deficient patients used higher homology than HD, specifically ATM$^{-/-}$, Ligase 4$^{-/-}$. These results align with previous work that observed ATM$^{-/-}$ mice using more MH in SM-SG1 joints than WT (*209*) and with the Ligase 4$^{-/-}$ human B cell line, Nalm-6. The latter showed that switch-joints were repaired using long stretches of homology promoted by the compensation of NHEJ with single-strand annealing (SSA), which uses more than 25 nt of homology to repair DNA breaks (*210*). On the other hand, results in murine B cell line CH12 showed a higher homology usage only in Rif1$^{-/-}$ CH12. This result agrees with the promotion of HR when Rif1 is silenced in Henrietta Lacks (HeLa) cells (*211*), likely because Rif1 directly inhibits HR by competing with BRCA1 for DNA foci binding (*104*, *105*).

Inversions in CSR were introduced in 2015 when they discovered that they rarely occurred in healthy samples, but upon cNHEJ protein depletion, they became more frequent (*178*). We found that inversions occurred at a median frequency of one per eff_ncluster in healthy humans and mice. However, Ligase 4$^{-/-}$ CH12 mean inversions per eff_nclusters was 25.8, unlike the Ligase 4$^{-/-}$ human sample, which showed no inversional event. It was shown that XRCC4$^{-/-}$ mice have increased inversional CSR joints in SM-SG1 and SM-SE than WT (*212*). XRCC4 is a protein that forms a heterodimer with Ligase 4 during cNHEJ to ligate the strands of a break. Our results could be explained by the fact that XRCC4/Ligase 4 heterodimer is necessary to limit the formation of inversional CSR joints. Inversions in human switch-joints have not been reported yet (*178*, *205*, *213*, *214*); therefore, discrepancies between our data and literature may result from species-specific features (*215*).

We introduced a set of novel features to characterize the switch-joint comprising the classification of breakpoints and implemented them in SWIBRID. DNA repair-deficient cells are prone to leave DNA breaks unrepaired and enter apoptosis (*216*). Therefore, it is expected to see fewer DNA breaks repaired in DNA repair deficiencies. Therefore, we novelly implemented single-, multiple- and within-breakpoints that differentiate between breakpoints that result in direct-join between SM and another switch region, sequential switching or intra-switch deletions, respectively. Analysis in human and mouse samples shows that multiple- and within-breakpoints are less frequent in DNA repair deficiencies than in HD or WT samples. Considering that productive CSR can result from a minimum of one single-breakpoint between SM and any other switch region, while multiple- and within-breakpoints are not essential to generate a productive CSR, it is likely that DNA repair deficient samples only acquired single-breakpoints. Potentially, DNA repair deficient samples that acquired multiple- and within-breakpoints could not repair them and eventually died, not being part of our analysis. Thus, DNA repair deficient B cells have a higher chance of surviving when they experience fewer DNA breaks.

In conclusion, SWIBRID is a tool that can support the research of CSR through the long-read sequencing and characterization of switch-joints. SWIBRID uses known features, namely, homology and inversions, and novel features, namely, eff_nclusters and classified breakpoints. SWIBRID is crucial to analyze samples

with different degrees of diversity since the standardization by eff_nclusters weights features among the samples equally.

8.3. Genomic J-CH1 inserts identified by SWIBRID mainly derive from introns, unlike RNA J-CH1 inserts that rely on splicing

J-CH1 inserts are found between the J segment and CH1 exons in the IGH locus. They have been thoroughly studied in Ig transcripts RNAs (*97*) and, less extensively, in genomic DNA (*94*). Pieper et al., 2017 studied J-CH1 inserts in genomic DNA using the same primers I used and sequencing with Illumina and MinION. In their study, the J-CH1 insert frequency is higher when using Illumina than MinION, 7 and 0.8 J-CH1 inserts per 1 000 reads, respectively. Long-amplicon sequencing by Illumina requires amplicon fragmentation and computational re-assembly, which can lead to both creation and deletion of clones, especially when the re-assembly is done in highly repetitive sequences as the switch regions. In my thesis, I used MinION to avoid the computational re-assembly of reads and improved the analysis, obtaining as many as 2 inserts per 1 000 reads.

Our study is the first to analyze J-CH1 inserts using genomic DNA in more than six HD and found a total of 31 880 unique J-CH1 inserts. SWIBRID J-CH1 inserts spanned introns, intergenic regions and exons, in order of frequency. From now on, I would refer to the J-CH1 insert origin characteristics as the J-CH1 insert signature. Interestingly, the J-CH1 insert origin signature (intron > intergenic > exon) resembled more accurately V-DJ inserts than J-CH1 inserts found in RNA, being introns the most frequently donated substrate for insertion. Also in agreement with Pieper et al., 2017 that found that J-CH1 inserts in the genomic DNA were enriched in introns (*94*). Based on the insert signatures, Lebedin et al., 2022 hypothesized that inserts derived from different genes depending on their incorporation site. Unlike Lebedin et al., 2022 I hypothesize that the source of the inserts incorporated into the IGH locus is mainly intronic, intergenic, and lower in exonic origin, regardless of the incorporation position.

My hypothesis is based on two facts. Firstly, the recombined VDJ is an exon; consequently, a genomic fragment inserted into the VDJ does not require independent splicing. As part of the VDJ domain, the inserts will be present in the transcripts and will be translated. In contrast, the J segment and the CH1 exon are separated by an intron (switch regions) spliced out upon transcription (*8*). Inserts within J and CH1 coding sequences (J-CH1 inserts) are thus expected to be spliced out unless they would carry matching splicing sites (donor and acceptor) (Figure *23*). It has been observed that cryptic splicing sites occur in as low as 9% of the cases in introns (*217*). Therefore, we can consider that the occurrence of matching splicing sites (donor and acceptor) in IGH inserts is odd; therefore, V-DJ inserts in RNA would resemble more the signature found in DNA than J-CH1 inserts in RNA, which would be biased by splicing. Pursuing this hypothesis should follow by sequencing VDJ recombined domains in the gDNA.

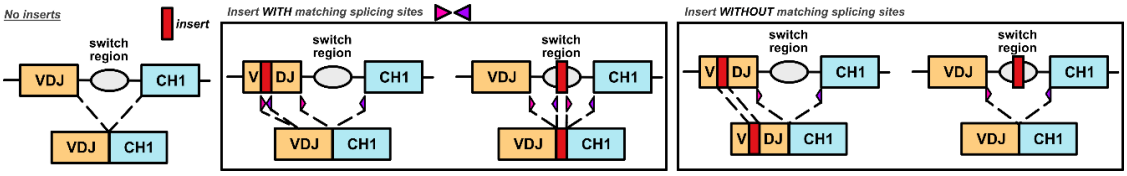


Figure 23. Splicing scenarios supports the insert origin.

Scheme showing splicing scenarios of an insert within V-DJ and J-CH1 with (middle) and without (right) matching splicing sites. Pink and purple triangles indicate the matching splicing sites

J-CH1 inserts are highly infrequent, and their origin is not yet understood (*93–95*, *97*). Thus, their appearance in Ig transcripts and between switch regions in the DNA can be confused with artifacts since no acquisition mechanism is known to date. J-CH1 inserts could result from a PCR amplification artifact by LongAmpTaq Pol since the most common artifact Taq polymerases produces is a chimeric amplicon (*218*). However, if J-CH1 inserts were artifacts, the switch-PCR would have to create a crossover product in which the DNA strand of the switch region would align to a DNA fragment, namely J-CH1 insert, and then the DNA fragment would need to align again to the switch region. This process is highly improbable to happen. Moreover, as observed in manual analysis, J-CH1 inserts do not always share homology with the flanking switch region sequences. Thus, I consider that J-CH1 inserts are not artifacts and should be considered to account for the diversification processes of the antibody.

Introns are the most significant donor source of V-DJ and SWIBRID J-CH1 inserts, and non-degraded splicing byproducts could explain such an enriched signature. Introns often are degraded upon splicing; however, they can also become stable and have secondary functions. For example, long-non-coding-RNAs regulate the alternative splicing of pre-mRNAs (*219*) or micro-RNAs silence genes to regulate their gene expression (*220*). However, no intronic origin of SWIBRID J-CH1 inserts matched significantly higher to long-non-coding- or micro-RNA databases than the control (*221–223*). Nonetheless, the source of J-CH1 inserts in the IGH locus could originate from late degraded splicing byproducts. Experiments targeting the degradation of introns after splicing, e.g., inhibition of the RNA debranching enzyme (DRB1)-dependent pathway to accumulate lariat intermediates after splicing (*224*), could be done to bring light to this hypothesis.

In conclusion, the study of J-CH1 insert origin using SWIBRID matched the previously published signature from V-DJ inserts found in transcripts and J-CH1 inserts found in genomic DNA – intronic > intergenic > exonic – suggesting that the latter is biased by splicing. Despite the originally intronic SWIBRID J-CH1 inserts did not match long-non-coding- or micro-RNAs, further research manipulating splicing *in vitro* should be performed.

8.4. SWIBRID characterization of CH12 knockout switch-joint features is in line with the literature

B cell includes VDJ recombination, which occurs early during B cell maturation, as well as CSR and SHM, which are processes activated after antigen recognition. Cell lines have limitations to study B cell diversification. CH12 is a mouse B cell lymphoma cell line that mainly switches to IgA and has been extensively used for the analysis of CSR (*76*), while Ramos is a human B cell lymphoma cell line that represents a model for the analysis of SHM (*225*, *226*). In this thesis, I used CH12 WT, 53BP1$^{-/-}$, BRCA1$^{-/-}$, Ligase 4$^{-/-}$ and Rif1$^{-/-}$ to study switch-joints generated during *in vitro* CSR.

53BP1 is a protein known to interact with Rif1 and to promote cNHEJ (*104*), but it has also been related to HR during G2 phase (*227*). 53BP1$^{-/-}$ CH12 has been previously observed to switch mainly to IgA and almost not to IgG1 or IgG2b (*205*, *228*). This agrees with the exclusive SM-SA switch-joints observed in our results in 53BP1$^{-/-}$ CH12. Also, it could explain the low amount of eff_nclusters found in 53BP1$^{-/-}$ CH12 since the lack of class-switching to other isotypes reduces the possibilities of generating a variety of B cell clones. Moreover, 53BP1$^{-/-}$ was shown to have increased breakpoints in the IGH locus (*228*); however, we observed a lower amount of intra-switch deletions (within-breakpoints). Unfortunately, a comparison to Dev et al., 2015 is hampered by the fact that the methodology used to calculate the breakpoints in 53BP1$^{-/-}$ CH12 was not explained, making the analysis comparison less veridic. Despite this exception, SWIBRID data could replicate the results previously observed in 53BP1$^{-/-}$ CH12.

BRCA1 is a protein involved in homology-directed DNA repair, which is not essential during CSR but helps to repair long-lasting CSR breaks during the S phase (*229*). We observed that switch-joint features in BRCA1$^{-/-}$ CH12 were similar to WT, except for the direct switch-joints between SM and another switch region, where BRCA1$^{-/-}$ CH12 had the highest frequency from all genotypes. A high amount of single-breakpoints indicates a higher diversity of break positions in the switch-joints. One assumption is that a large number of breaks would give a higher spread value in SM and SA. Spread indicates the distance between the most upstream and downstream breakpoints in a switch region. However, SM and SA spread in BRCA1$^{-/-}$ CH12 was not significantly different from WT CH12. There was only a slightly higher tendency of BRCA1$^{-/-}$ CH12 SM spread than WT CH12. Interestingly, lack of BRCA1 was observed not to vary the number of switch-joints (*230*), agreeing with our results where we can observe that this deficiency and WT have similar eff_nclusters.

Ligase 4 is the enzyme that, in complex with XRCC4, end-ligates two dsDNA break ends and finalizes cNHEJ (*136*). As previously outlined in section 8.2, Ligase 4$^{-/-}$ CH12 switch-joints presented a high occurrence of inversions. Also, Ligase 4$^{-/-}$ CH12 showed a high SG3 spread. I have not found a factual explanation for the high spread in SG3. However, dsDNA break could be processed by aEJ (*188*) due to Ligase 4 loss, suffering resection in both directions in an attempt to find homology and repair the DNA break. SM and SA could experience a lower resection because homology between them can be found

more easily; however, SGs are less homologous to SM than SA, potentially increasing resection. Ligase $4^{-/-}$ CH12 results were aligned with published results.

Rif1 protects dsDNA breaks from resection during DNA repair and replication fork stalling (*104*, *231*). Rif1 loss impairs cNHEJ, the preferred DNA repair mechanism during CSR and promotes homology-directed repair (*104*, *232*). Our results show that $Rif1^{-/-}$ CH12 uses the highest 4-mer homology of all samples, which can be translated as that homology-directed DNA repair has repaired the breaks in $Rif1^{-/-}$ CH12 knockout. $Rif1^{-/-}$ CH12 had a very low spread in all switch regions and the lowest amount of eff_nclusters (2-fold less than the second lower, $53BP1^{-/-}$ CH12). These features depend on each other since every new eff_ncluster provides new breakpoint information for the spread unit since the more clusters a sample has, the more diverse breakpoints can contribute to the spread unit. However, it is worth noting that the low spread unit indicates that the breakpoints occur in a very narrow region of the switch regions. The low number of eff_nclusters could be explained by the lack of cNHEJ. It is important to note that all of the $Rif1^{-/-}$ CH12 replicates gave rise to many reads per cluster, which could mean that upon activation, most of the $Rif1^{-/-}$ clones died, and the few successfully class-switched B cells, proliferated and occupied the majority of the culture.

Resection is a characteristic that helps understand the DNA repair pathway involved in a repair. Resection can help identify 53BP1 (*233*), Rif1 (*105*) and BRCA1 (*234*) knockouts, considering that the two first inhibit resection while the latter promotes it. Resection is measured by quantifying the length of DNA missing from a fixed DNA break. Considering that we manage a set of diverse switch-joints with random breakpoints, SWIBRID could not measure resection. Nonetheless, resection was not required to separate the genotypes in the PCA; however, the analysis could benefit from having such a feature. SWIBRID could adopt the method that Vincendeau et al., 2022 followed to measure resection by measuring the length of the amplicons. Longer amplicons would relate to less resection and vice versa.

In conclusion, CH12 knockout switch-joint features analyzed by SWIBRID agree with the literature, validating our results. New features of the genotypes, such as higher diversity in SM direct joins to other switch regions in $BRCA1^{-/-}$ CH12 and low amount of eff_nclusters in $Rif1^{-/-}$ CH12, show how SWIBRID contributes to the analysis of DNA repair deficiencies by novel features. SWIBRID can open a new way to analyze DNA repair during CSR discovering new features.

8.5. Characteristics of switch-joint derived from DNA repair-deficient patients are in line with published data

Analysis of DNA repair can be performed using inactive or absent proteins. However, the outcomes of both experiments are, in some cases, very different. For instance, mice carrying inactivated ataxia telangiectasia and Rad3-related protein (ATR) ($ATR^{+/KD}$) or loss of ATR ($ATR^{+/loss}$) experience different outcomes. $ATR^{+/KD}$, but not $ATR^{+/loss}$, caused ssDNA defects causing, for instance, male

infertility. The reason was that the kinase function of ATR is necessary for its mobility within the DNA break. However, inactivation and immobility caused an accumulation of ATR/RPA foci in the break, impairing its repair (*235*). This example resembles the importance of studying the impact of DNA repair proteins inactivating them and also as a loss of it. Since I analyzed the inhibition of proteins to investigate the protein function in J-CH1 insert incorporation, I also studied human patients with deficiencies in BRCA1$^{-/-}$, Ligase 4$^{-/-}$, ATM$^{-/-}$, NIPBL$^{-/-}$ and AID$^{-/-}$ to understand the physical role of the proteins in J-CH1 incorporation.

First, I researched their *in vivo* acquisition of J-CH1 inserts. BRCA1$^{-/-}$, Ligase 4$^{-/-}$, ATM$^{-/-}$, and AID$^{-/-}$ human patients lacked J-CH1 inserts. Even though these three proteins do not have a common background (*7*, *104*, *136*, *236*), their absence caused the same outcome. However, the NIPBL$^{-/-}$ patient had a slightly higher diversity of J-CH1 inserts than HD. NIPBL is responsible for cohesin loading into the DNA. Cohesin is a protein complex that connects sister chromatids, and that has also been shown to play a role in DNA repair, e.g., facilitates the homology search during HR DNA repair (*237*). NIPBL impairment has been observed to stimulate aEJ, which leads to a higher chance of translocation between chromosomes (*176*). Since aEJ induces intrachromosomal aberrations, I hypothesize that stimulation of aEJ could promote J-CH1 inserts incorporation.

Second, I investigated the characteristics of their genomic scars in switch-joints. In this case, I aimed at differentiating DRdef patients from HD using the features of homology used for repair, classification of breakpoints and inversional events by PCA. Ligase 4$^{-/-}$, ATM$^{-/-}$, and AID$^{-/-}$ human patients differentiated from HD; however, BRCA1$^{-/-}$ and NIPBL$^{-/-}$ did not. I concluded that the latter two had breakpoint profiles similar to HD and could not be separated via PCA. I hypothesize that the reason is that BRCA1 absence would impair homology-directed DNA repair, promoting error-prone DNA repair mechanisms, e.g., cNHEJ (*238*). Since cNHEJ is the preferred DNA repair mechanism during CSR, we could expect not to observe differences between HD and BRCA1$^{-/-}$ breakpoint profiles. Following the same lines, NIPBL is part of the cohesin protein complex, which has been suggested to have the role of strand screening aiding RAD51 during HR repair (*237*). Thus, error-prone DNA repair can be expected in the absence of NIBPL and again, no differences in breakpoint profile between NIPBL$^{-/-}$ and an HD would be expected.

Third, I decided to deepen into the characteristics of the DNA repair patients with the highest affection in CSR: Ligase 4$^{-/-}$, ATM$^{-/-}$, and AID$^{-/-}$. The AID$^{-/-}$ donor had a very poor number of breakpoints. Considering that AID is the CSR-promoting enzyme and I analyzed switch-joints, I did not expect to find any breakpoint events. Studies in AID$^{-/-}$ mice showed that these mice present only high levels of IgM in serum (*239*). However, SM-SA1 and SM-SA2 joints could be observed in the human AID$^{-/-}$ sample. Two independent studies showed in mice that extensive transcription can cause spontaneous dsDNA breaks (*240*, *241*), suggesting that AID$^{-/-}$ switch-joints could have occurred as a random occurrence of DNA breaks due to the excessive transcription of switch regions during GLT production.

ATM$^{-/-}$ patients were composed of two biological replicates carrying different mutations. Nonetheless, both ATM$^{-/-}$ donors presented breakpoint profiles with a preference for SM-SA1 and SM-SA2 switch-joints. ATM is a protein that senses dsDNA breaks and phosphorylates proteins to promote both HR and cNHEJ (*242*, *243*), so theoretically, both mechanisms would be expected to be impaired. Nonetheless, it was described that ATM$^{-/-}$ mice handle dsDNA breaks using homology-driven repair during CSR (*209*). SM and SA regions share high sequence homology, as seen in the homology plots in section 7.4.3. Therefore, if ATM$^{-/-}$ handles the dsDNA breaks using homology-directed repair, we could expect to find SM-SA1 and SM-SA2 switch-joints. dsDNA breaks in SG regions would likely be left unrepaired, directing the cells to apoptosis. However, the absolute lack of SM-SGs joints also suggests that there could be a deficiency in cytokines promoting IgG switching.

Ligase 4$^{-/-}$ donor presented no switch-joints between SM-SG2 and SM-SG4, which aligns with the literature (*136*). Ligase 4 is part of the complex that joins the two strands from the DNA break during cNHEJ. Interestingly, Ligase 4$^{-/-}$ switch-joints pattern resembles the type of CSR occurring in children, which uses only the first part of the IGH locus (*207*). Even though we lack information about their age, we can assume that the Ligase 4$^{-/-}$ donor was a child because the life expectancy of this deficiency is low (*244*), and it would agree with the breakpoint profile.

In conclusion, the data obtained using SWIBRID in human DNA repair deficient patients aligns with the literature. BRCA1$^{-/-}$ and NIPBL$^{-/-}$ human patients did not present an impairment in the CSR of B cells, unlike Ligase 4$^{-/-}$, ATM$^{-/-}$, and AID$^{-/-}$ human patients. Ligase 4$^{-/-}$, ATM$^{-/-}$, and AID$^{-/-}$ had a preference for SM-SA1 and SM-SA2 joints and a lack of SM-SGs joints, likely due to the usage of homology-driven repair

8.6. SWIBRID identifies unique CSR events in B cells: switch-joint versus VDJ B cell repertoire

Analysis B cell repertoires using sequencing of VDJ recombined domains and switch-joints gave rise to vastly different values. Specifically, switch-joint clustering done by SWIBRID elicited 3.2- and 8.5-fold higher SWIBRID-clones than VDJ-sequencing in the analysis of 100 000 and 50 000 cells, respectively.

I expected to obtain up to 2-fold higher number of SWIBRID-clones than VDJ-clones. This expectation was based on the fact that i) one B cell can carry two switched alleles and ii) the occurrence of non-sense-mediated decay for non-productively recombined Ig transcripts (*245*, *246*). Firstly, B cells contain two alleles with two independent IGH loci. During BM diversification, only the IGH in one allele undergoes VDJ recombination. If the rearrangement does not result in a functional antibody or it gives rise to an autoimmune antibody, the B cell will allow the allelic inclusion and perform the VDJ recombination in the other allele. Since allelic exclusion acts via epigenetics to avoid gene rearrangement (*92*), I hypothesized that a newly included allele would also allow the CSR and SHM

events to occur. In that case, a cell would carry two class-switched IGH. However, despite carrying two class-switched IGHs, only one would be transcribed due to allelic exclusion (*91*). In this case, SWIBRID would amplify both alleles and consider one unique B cell as two unique B cell clones, while VDJ-sequencing would only consider the presence of one unique B cell based on the CDR3 transcribed.

Secondly, non-sense mutations in the exons of the IGH lead to premature stop codons. Cells have a mechanism to degrade the mRNA-carrying stop codons in the middle of the sequence and avoid wrongly folding proteins that potentially cause damage. This mechanism is called non-sense-mediated decay (NMD). NMD in B cells degrades 80% of non-sense Ig transcripts in active cells (memory B cells) but only 50% of non-sense Ig transcripts in resting B cells (naïve B cells) (*247*). Considering that we are studying memory B cells, we can expect that if B cells would carry a non-sense Ig transcript, only 20% would be available for VDJ-sequencing, unlike in SWIBRID, which sequences the DNA not subjected to NMD.

Another explanation for the high number of SWIBRID clones could be the deletional events in the switch regions besides CSR. SWIBRID identifies a clone as the unique switch-joint in a library of B cells. It considers CSR events, such as the joint between two different switch regions, but also non-CSR-related events, such as the J-CH1 inserts and intra-switch region deletions. The switch-joint is the genomic scar that occurs in the introns of the antibodies, while VDJ segments are exons. Cells carrying mutations that can change the folding of proteins undergo apoptosis triggered by endoplasmic reticulum stress (*248*). These mutations mainly involve those happening in exons and, for instance, create cryptic splicing sites, amino acid exchanges or stop codons. B cells with deleterious mutations in the VDJ likely undergo apoptosis or gene editing, while the same mutations occurring in the switch-joints would not be a subject of selection. Thus, we expect to see more switch-joint mutations and, therefore, more SWIBRID-clones.

VDJ- and SWIBRID-repertoire analysis results show that B cells carrying the same V-D-J-C domains undergo different diversification processes. SWIBRID-clones represent the degree of diversification that a B cell population has gone through, and although it maintains a direct relationship with the amount of VDJ-clones, it differs among donors. Division of SWIBRID- versus VDJ-clones within the same donor, SWIBRID/VDJ-clone ratio, compares how frequently B cells underwent CSR in a donor. SWIBRID/VDJ-clone ratio in 50 000 cells was the highest in HD1 with 16 and the lowest in HD2 with 7. Differences in the SWIBRID/VDJ-clone ratio suggest that people can bare different amounts of DNA breakage or that their DNA repair is differently efficient.

In conclusion, SWIBRID and VDJ quantification of the B cell repertoire identify different clones by definition. SWIBRID observed more unique B cell clones, suggesting that different SWIBRID unique B cell clones carry the same recombined VDJ and isotype. Comparing the SWIBRID/VDJ-clone ratio between

three HD suggests that people have diverse DNA breakage burdens or DNA repair efficiency.

8.7. Inhibition of DNA repair proteins in *in vitro* activated human B cells impairs B cell diversification and provides insights of the J-CH1 insert source

The experiments inhibiting Ligase 4, PARP1 and ATM using SCR7, Olaparib and KU-60019 in primary human B cells were performed to gain insights into the J-CH1 insert mechanism. Even though most of the inhibitors were highly toxic to the B cells in culture, they revealed hints about the J-CH1 insert source.

Inhibition of Ligase 4 via SCR7 gave rise to a very low amount of eff_nclusters, which reflects the impairment of cNHEJ, the preferred DNA repair mechanism during CSR (*136*). SCR7 treatment was highly toxic for the primary human B cells; in fact, only one out of four replicates generated switch-joints. The only replicate that elicited enough switch-joints had more than 20 unique J-CH1 inserts, which was higher than the control. The origin of these J-CH1 inserts was closer to ERFS than the control, which matches the origin of J-CH1 inserts found in memory B cells under the classification of Lebedin et al., 2022. Considering the rate of switch-joint formation failure upon SCR7 treatment, I decided not to pursue this inhibitor experiment. Indeed, generating three viable replicates would have involved performing at least twelve SCR7 inhibitor experiments. The data from this replicate suggest that Ligase 4 inhibition increases J-CH1 insert incorporation.

Inhibition of PARP1 via Olaparib gave rise to a higher amount of eff_nclusters than the DMSO control. However, studies in rhabdomyosarcoma and ovarian cancer cell lines showed that Olaparib at 1.5 µM and 10 µM either did not interfere with or decrease the proliferation of the cells (*249*, *250*). Proliferation measures were done in both cases, counting the number of cells. Even though, diversity quantified by SWIBRID does not have to equal proliferation, if there were a lower proliferation, less diversity is to be expected. Thus, I hypothesize that we observed more eff_nclusters than the literature because PARP1 inhibition benefits CSR. This hypothesis is based on the fact that *in vitro* activation of primary human B cells in the presence of 1.25 µM Olaparib would impair aEJ, which is compensated by using cNHEJ, the preferred DNA repair during CSR (*251*). Inhibition of PARP1 also elicited an extremely low J-CH1 insert diversity. Since PARP1 is part of the aEJ, this could hint at the involvement of aEJ in incorporating J-CH1 inserts (*164*). In 2013, aEJ was linked to the occurrence of translocations. They demonstrated that the IGH locus translocates to other chromosomes using fluorescence *in situ* hybridization (FISH) (*176*). Since aEJ has been linked to intrachromosomal genomic aberrations, it may also promote the incorporation of J-CH1 inserts in the IGH.

On the other hand, we modulated the replication DNA machinery and promoted the replication fork stall via HU. The replication fork stalling has been linked to the breakage of ERFS (*162*), a group of genomic sites found closer to J-CH1

insert origins in antibody transcripts (*97*). HU treatment gave rise to a highly variable J-CH1 insert frequency between the donors, originating more proximal to telomeres and ERFS than DMSO control. The inconsistency between the J-CH1 insert frequency among different donors prevents us from concluding ERFS breakage and J-CH1 insert. However, the data suggests that the breakage of ERFS by HU increases the J-CH1 insert source closer to telomeres and ERFS.

Finally, inhibition of ATM via KU-60019 at different dosages decreased the diversity excessively. It was already observed that $ATM^{-/-}$ mice contain up to 120-fold fewer SM-SG1 and SM-SE switch-joints than WT (*209*). Having fewer switch-joints translates to eliciting a lower number of clusters. Also, the different dosages of KU-60019 elicited 5-fold fewer unique J-CH1 inserts than the DMSO control, which made it challenging to analyze their origin.

In conclusion, SWIBRID recorded data supporting ERFS could be a source of J-CH1 inserts. Also, some hints point towards the involvement of aEJ in J-CH1 inserts incorporation, which will be discussed below. Unfortunately, inhibiting these proteins decreases the diversity so much that little data was obtained. Therefore, I decided to focus more on the experiments that provided hints about the mechanism of J-CH1 inserts acquisition, the analysis of TERT inhibition.

8.8. Inhibition of TERT decreases the amount of J-CH1 inserts within SM

Insert classification based on transcripts (*97*), divided J-CH1 inserts into naive J-CH1 inserts, with origins very close to telomeres, and memory J-CH1 inserts, with origins also close to telomeres, but especially to ERFS (*97*). This division raised the hypothesis that the J-CH1 insert incorporation could be related to telomere processing, especially the naïve J-CH1 inserts. AID was observed to produce off-target deaminations in telomeres due to the presence of WRCY motifs (*252*). Telomeres and switch regions share similarities in their sequences. They are both repetitive C-rich sequences, which sequences are transcribed to produce non-translatable transcripts that have the potential to form G-quadruplex structures (*252*, *253*). The resemblance is so high that AID has been observed to off-target the telomeres in a process that shorten the telomeres (*252*, *254*). Therefore, I hypothesize that TERT also off-target switch regions and could play a role in the J-CH1 insert incorporation. Specifically and after finding telomeric repeats as J-CH1 inserts, TERT could synthesize *de novo* the fragments that result as J-CH1 inserts.

Using different doses, I inhibited TERT using BIBR1532 during the activation process of human naïve B cells. The inhibition of TERT did not affect the incorporation of J-CH1 inserts nor the specific incorporation of J-CH1 inserts constituted of telomeric repeats. Lebedin et al., 2022 showed that J-CH1 insert found in naïve B cells originated from genes closer to telomeres than those found in memory B cells. Naïve B cells have an intact constant region in their IGH. Thus, Lebedin et al., 2022 analyzed J-CH1 inserts in naïve B cells incorporated between the J-segment and the CH1 exon of IGHM, potentially incorporated into the SM. Following these lines, I analyzed the J-CH1 inserts that were potentially

incorporated before CSR occurred, and those would be flanked by the SM sequence. In order to do that, I classified the J-CH1 inserts that were found to be embedded within a switch region into the different isotypes. The results show that the J-CH1 inserts incorporated flanked by SM sequence, and not any other isotype, decreased their frequency in a BIRB1532 dose-dependent manner. These results indicate that TERT may have a role in incorporating J-CH1 inserts within SM, which could happen in naive or memory B cells.

In conclusion, TERT inhibition did not impair J-CH1 inserts in *in vitro* activated primary human B cells. Nonetheless, the data suggest that TERT is involved in the acquisition of J-CH1 inserts in the SM region.

8.9. *In vitro* and *in vivo* analysis of J-CH1 inserts incorporation hints towards the implication of aEJ

Results on J-CH1 inserts acquisition obtained by SWIBRID suggest that aEJ is involved in their incorporation into the IGH. Specifically, the modulation of aEJ via NIPBL$^{-/-}$, PARP1 or Ligase 4 inhibition hinted towards the implication of aEJ in incorporating J-CH1 inserts. NIPBL$^{-/-}$ has been observed to promote aEJ because it was found that the genomic scars occurring in this deficiency were rich in the usage of MH < 4nt and resection and had a very low amount of blunt DNA repairs (*176*). The NIPBL$^{-/-}$ human sample elicited a slightly higher J-CH1 insert frequency than controls which could be linked to the intrachromosomal aberrations from the aEJ. PARP1 inhibition impairs aEJ since PARP1 works at the early stages of the DNA repair mechanism facilitating the DNA synapsis (*164*). Results show that PARP1 inhibition elicited a very low J-CH1 insert frequency which could be a consequence of the inhibition of aEJ. Finally, it has been observed that catalytically inactive Ligase 4, which could resemble Ligase 4 inhibition, promotes aEJ via Ligase 3 (*188*). In our results, Ligase 4 inhibition elicited a relatively high J-CH1 insert frequency, which could be linked to the promotion of aEJ.

aEJ is a DNA repair pathway that is error-prone and uses <6 nt MH for dsDNA break repair. Unfortunately, the analysis of J-CH1 insert homology in the switch-joints incorporation could not be performed because the MinION error rate is 5%. Analysis of 5 nt regions in the sequence entails a risk of analysis of MinION errors. In order to analyze the homology used for repair and considering that the average length of a clone is 2.5 Kb, we would need at least 20 reads from the same switch-joint and create a consensus sequence out of them to have a low chance of finding mutations in the joint region.

In conclusion, analysis of human samples and *in vitro* inhibition of proteins hinted towards the implication of aEJ to the J-CH1 insert incorporation. The research should follow with experiments modulating aEJ in different ways, for example, inhibiting Ku70/80 or Ligase III to promote or inhibit aEJ, respectively (*114*).

8.10. Future directions in the study of J-CH1 inserts in genomic DNA

The analysis of J-CH1 insert incorporation in genomic DNA allowed the analysis of more than 31 000 inserts. However, the mean length of the SWIBRID J-CH1 inserts was around 10-fold less than expected from published data (*97*). For instance, Lebedin et al., 2022, identified an insert originating in the casein kinase I isoform delta (*CSNK1D*) gene consisting of two exons separated originally by a 10 kb intron. They could amplify it because the analysis was done in transcripts, and the exons would be spliced. However, the chance of finding that insert in genomic DNA is almost impossible because the current methodology to obtain switch-joints is a PCR limited to 5kb. This limitation underestimated the analysis and real frequency of J-CH1 inserts. Therefore, attempts to optimize the study of the J-CH1 inserts should be addressed. In order to do that, the amplification of switch regions using gDNA could be improved.

The switch-joint PCR could be improved by placing the SG RV and SA RV primers in the exons rather than switch regions (*205*). Unlike switch regions, the exons are not as targeted during B cell diversification, as observed in the single nucleotide variant plots in Machado et al., 2022. Thus, exons will always be present in the IGH locus and using them to place primers would ensure that even when the break occurs very close to the exon, a sustainable amount of amplicons can still be amplified. In this approach, we must consider that the PCR would need to amplify a longer stretch of DNA.

Another approach would use the suppression PCR for switch-joints. The suppression PCR is a method that biases the PCR amplification of longer amplicons by using primers designed to form hairpins when in short amplicons that would impair their amplification. The suppression PCR was developed by Lukyanov et al., 1995 and was used by Lebedin et al., 2022 to enrich the amplification of inserts in antibody transcripts (*97*, *143*). The suppression PCR would be optimal for the analysis of insert origin since it would capture a higher amount of inserts in the IGH locus. However, suppression PCR limits the quantification of B cell clones since it biases the PCR amplification to larger amplicons (*97*).

A larger-scale approach would amplify the switch regions using Linear AMplification High-Throughput Genome-Wide Translocation Sequencing (LAM-HTGTS) (*255*). LAM-HTGTS uses only one primer, which could bind upstream to SM, to amplify a stretch of DNA limited by the elongation time in the PCR cycler and the number of cycles. This approach is the most unbiased PCR approach to date and would allow identifying large inserts at the DNA level. LAM-HTGTS could also amplify VDJ inserts using a primer amplifying upstream from SM to confirm that the insert signature in both J-CH1 and VDJ inserts are identical.

And finally, the approach that could be used is the non-PCR amplified analysis of inserts in the IGH locus. Although LAM-HTGTS offers an opportunity to amplify PCR products without a size bias, artifacts could arise from chimera formation (*255*, *256*). Therefore, we could also opt for a non-PCR-based amplification of the switch-joints using Cas9 fragment enrichment (*257*). Cas9 can precisely cut, using one or two CRISPR RNAs, the extremes of the switch-joints. Then, the fragments can be directly prepared for MinION loading. However, not all the

alleles in the cells of interest would contain switch-joints, even if they are memory B cells. This would be the case for cells that only have one switched allele. When the IGH locus in one allele is not switched, it can give rise to fragments as big as 268 Kb (SM-SA2) that should be filtered out before loading them to MinION in order only to consider switch-joints.

In conclusion, improving the switch-joint amplification will ensure a better J-CH1 insert analysis to obtain hints about the J-CH1 insert incorporation. Different methods that could unbias the approach and enable the identification of larger inserts are readily available and could be used to optimize the method.

8.11. The future of SWIBRID: The DNA repair biomarker envisaged as a diagnostic tool to predict disease

SWIBRID, or the DNA repair biomarker, can become a diagnostic tool used in the clinic to predict diseases. The diseases it will be able to predict would be those which cause lies in DNA repair or B cell biology. Thus, cancer, neurodevelopmental diseases, autoimmune diseases or immunodeficiencies would be the main target of the DNA repair biomarker. Cohorts of different diseases with age- and gender-matching healthy controls would allow a patient sample to be tested against several diseases (Figure *24*). These cohorts might be challenging to obtain because i) diseases are rare, such as CVID, which occurs in 1:25 000 people (*196*) or ii) the life expectancy of diseases is low, like Cockayne syndrome, whose patients live up to 16 years old (*258*), parents could restrict access.

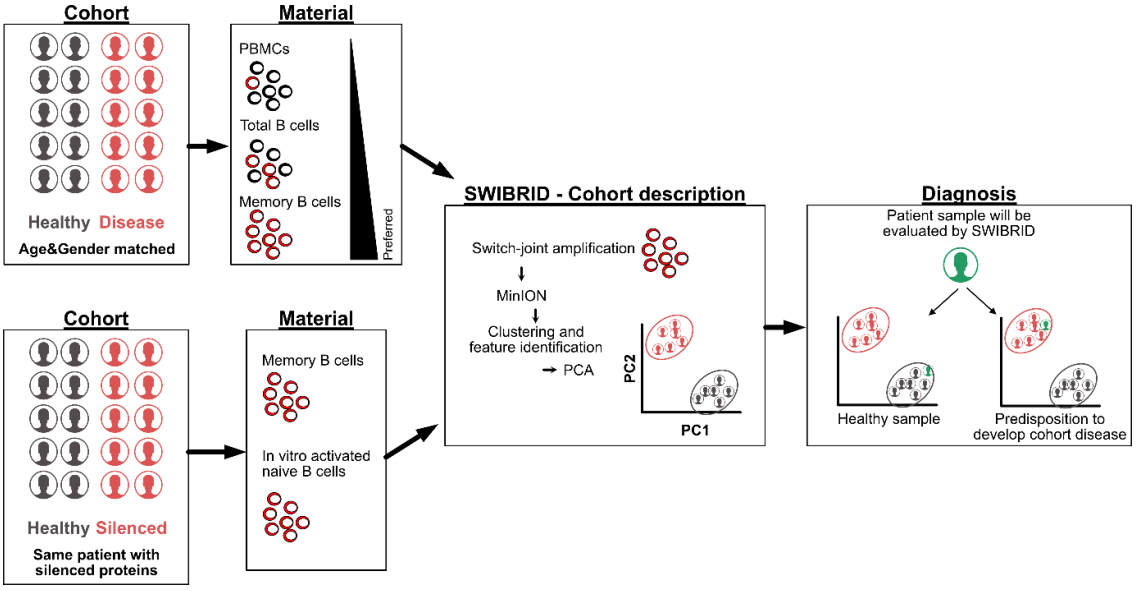


Figure 24. DNA repair biomarker scheme.

Cohort patients in grey represent the healthy controls, and patients in red represent the disease or silenced donors. Cells represented in red stand for cells of interest. PCA=principal component analysis.

Therefore, cohorts could also be accomplished by silencing specific proteins in primary human B cells of HD (Figure *24*). The idea is to divide every HD material into two groups: i) the primary human B cells would be activated as such, and ii) the primary human B cells would undergo silencing of target proteins, e.g., ATM, and subsequentially would be activated. This approach would generate a group of sick and healthy matching samples for protein deficiencies that are rare, and thus, their material can be challenging to obtain (*259*). It will be essential to consider that the silenced B cells would be activated *in vitro* and not *in vivo*. Considering that *in vitro* and *in vivo* activation do not elicit the same CSR outcomes (*97*, *260*), comparing *in vitro* activation with the *in vivo* switch-joint of the donors using their $CD27^{+}IgG/A^{+}IgM^{-}$ B cells should be performed. The healthy control group must be composed exclusively of non-carriers of DNA repair protein mutations. Addressing this issue would involve the screening of such mutations either by SNP screening or, preferably, exome sequencing.

In the future, we will analyze new patient switch-joints using SWIBRID and identify a person's risk of having a particular protein deficiency. Considering that SWIBRID would count with samples with numerous DNA repair protein deficiencies, we could analyze the risk of several protein deficiencies using the same data from SWIBRID.

In conclusion, efforts focus on developing the DNA repair biomarker using SWIBRID. We envisage this tool to diagnose the predisposition to develop diseases by estimating the efficiency of DNA repair in switch-joints. The DNA repair biomarker may improve how DNA repair malfunction is analyzed in the clinics and help understand a person's risk of developing diseases like cancer.

9. LITERATURE

1. D. L, M. E., *Janeway's immunobiology* (2016).

2. I. Mikocziova, V. Greiff, L. M. Sollid, Immunoglobulin germline gene variation and its impact on human disease. *Genes Immun*. 22, 205–217 (2021).

3. A. R. Rees, Understanding the human antibody repertoire. *Mabs*. 12, 1729683 (2020).

4. Y. Elhanati, Z. Sethna, Q. Marcou, C. G. Callan, T. Mora, A. M. Walczak, Inferring processes underlying B-cell repertoire diversity. *Philosophical Transactions Royal Soc B Biological Sci*. 370, 20140243 (2015).

5. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes and T-cell recognition. *Nature*. 334, 395–402 (1988).

6. H. Ru, M. G. Chambers, T.-M. Fu, A. B. Tong, M. Liao, H. Wu, Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-RAG2 Complex Structures. *Cell*. 163, 1138–1152 (2015).

7. M. Muramatsu, K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, T. Honjo, Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. *Cell*. 102, 553–563 (2000).

8. D. D. Dudley, J. Chaudhuri, C. H. Bassing, F. W. Alt, Mechanism and Control of V(D)J Recombination versus Class Switch Recombination: Similarities and Differences. *Adv Immunol*. 86, 43–112 (2005).

9. A. M. Müller, A. Medvinsky, J. Strouboulis, F. Grosveld, E. Dzierzakt, Development of hematopoietic stem cell activity in the mouse embryo. *Immunity*. 1, 291–301 (1994).

10. Y. H. Wang, J. Nomura, O. M. Faye-Petersen, M. D. Cooper, Surrogate light chain production during B cell differentiation: differential intracellular versus cell surface expression. *J Immunol Baltim Md 1950*. 161, 1132–9 (1998).

11. K. Pieper, B. Grimbacher, H. Eibel, B-cell biology and development. *J Allergy Clin Immun*. 131, 959–971 (2013).

12. I. Isnardi, Y.-S. Ng, L. Menard, G. Meyers, D. Saadoun, I. Srdanovic, J. Samuels, J. Berman, J. H. Buckner, C. Cunningham-Rundles, E. Meffre, Complement receptor 2/CD21− human naive B cells contain mostly autoreactive unresponsive clones. *Blood*. 115, 5026–5036 (2010).

13. H. Murai, H. Hara, T. Hatae, T. Kobayashi, T. Watanabe, Expression of CD23 in the germinal center of thymus from myasthenia gravis patients. *J Neuroimmunol*. 76, 61–69 (1997).

14. W. Dechkhajorn, S. Benjathummarak, S. Glaharn, U. Chaisri, P. Viriyavejakul, Y. Maneerat, The activation of BAFF/APRIL system in spleen and lymph nodes of Plasmodium falciparum infected patients. *Sci Rep-uk*. 10, 3865 (2020).

15. A. Enders, A. Short, L. A. Miosge, H. Bergmann, Y. Sontani, E. M. Bertram, B. Whittle, B. Balakishnan, K. Yoshida, G. Sjollema, M. A. Field, T. D. Andrews, H. Hagiwara, C. C. Goodnow, Zinc-finger protein ZFP318 is essential for expression of IgD, the alternatively spliced Igh product made by mature B lymphocytes. *Proc National Acad Sci*. 111, 4513–4518 (2014).

16. P. D. Pioli, I. Debnath, J. J. Weis, J. H. Weis, Zfp318 Regulates IgD Expression by Abrogating Transcription Termination within the Ighm/Ighd Locus. *J Immunol*. 193, 2546–2553 (2014).

17. S. Pillai, A. Cariappa, The follicular versus marginal zone B lymphocyte cell fate decision. *Nat Rev Immunol*. 9, 767–777 (2009).

18. J. P. Pereira, L. M. Kelly, Y. Xu, J. G. Cyster, EBI2 mediates B cell segregation between the outer and centre follicle. *Nature*. 460, 1122–1126 (2009).

19. M. A. Linterman, How T follicular helper cells and the germinal centre response change with age. *Immunol Cell Biol*. 92, 72–79 (2014).

20. T. Okada, M. J. Miller, I. Parker, M. F. Krummel, M. Neighbors, S. B. Hartley, A. O'Garra, M. D. Cahalan, J. G. Cyster, Antigen-Engaged B Cells Undergo Chemotaxis toward the T Zone and Form Motile Conjugates with Helper T Cells. *Plos Biol*. 3, e150 (2005).

21. K. G. C. Smith, T. D. Hewitson, G. J. V. Nossal, D. M. Tarlinton, The phenotype and fate of the antibody-forming cells of the splenic foci. *Eur. J. Immunol.* 26, 444–448 (1996).

22. D. Paus, T. G. Phan, T. D. Chan, S. Gardam, A. Basten, R. Brink, Antigen recognition strength regulates the choice between extrafollicular plasma cell and germinal center B cell differentiation. *J Exp Medicine*. 203, 1081–1091 (2006).

23. N. S. D. Silva, U. Klein, Dynamics of B cells in germinal centres. *Nat Rev Immunol*. 15, 137–148 (2015).

24. L. Krzyzak, C. Seitz, A. Urbat, S. Hutzler, C. Ostalecki, J. Gläsner, A. Hiergeist, A. Gessner, T. H. Winkler, A. Steinkasserer, L. Nitschke, CD83 Modulates B Cell Activation and Germinal Center Responses. *J Immunol*. 196, 3581–3594 (2016).

25. G. D. Victora, M. C. Nussenzweig, Germinal Centers. *Annu Rev Immunol*. 40, 1–30 (2022).

26. G. Koopman, H. K. Parmentier, H. J. Schuurman, W. Newman, C. J. Meijer, S. T. Pals, Adhesion of human B cells to follicular dendritic cells involves both the lymphocyte function-associated antigen 1/intercellular adhesion molecule 1 and

very late antigen 4/vascular cell adhesion molecule 1 pathways. *J Exp Medicine*. 173, 1297–1304 (1991).

27. R. Nakagawa, D. P. Calado, Positive Selection in the Light Zone of Germinal Centers. *Front Immunol*. 12, 661678 (2021).

28. M. Reynes, J. P. Aubert, J. H. Cohen, J. Audouin, V. Tricottet, J. Diebold, M. D. Kazatchkine, Human follicular dendritic cells express CR1, CR2, and CR3 complement receptor antigens. *J Immunol Baltim Md 1950*. 135, 2687–94 (1985).

29. D. P. Calado, Y. Sasaki, S. A. Godinho, A. Pellerin, K. Köchert, B. P. Sleckman, I. M. de Alborán, M. Janz, S. Rodig, K. Rajewsky, The cell-cycle regulator c-Myc is essential for the formation and maintenance of germinal centers. *Nat Immunol*. 13, 1092–1100 (2012).

30. Z. Shulman, A. D. Gitlin, J. S. Weinstein, B. Lainez, E. Esplugues, R. A. Flavell, J. E. Craft, M. C. Nussenzweig, Dynamic signaling by T follicular helper cells during germinal center B cell selection. *Science*. 345, 1058–1062 (2014).

31. A. B. Holmes, C. Corinaldesi, Q. Shen, R. Kumar, N. Compagno, Z. Wang, M. Nitzan, E. Grunstein, L. Pasqualucci, R. Dalla-Favera, K. Basso, Single-cell analysis of germinal-center B cells informs on lymphoma cell of origin and outcome. *J Exp Med*. 217, e20200483 (2020).

32. D. Suan, N. J. Kräutler, J. L. V. Maag, D. Butt, K. Bourne, J. R. Hermes, D. T. Avery, C. Young, A. Statham, M. Elliott, M. E. Dinger, A. Basten, S. G. Tangye, R. Brink, CCR6 Defines Memory B Cell Precursors in Mouse and Human Germinal Centers, Revealing Light-Zone Location and Predominant Low Antigen Affinity. *Immunity*. 47, 1142-1153.e4 (2017).

33. R. Shinnakasu, T. Inoue, K. Kometani, S. Moriyama, Y. Adachi, M. Nakayama, Y. Takahashi, H. Fukuyama, T. Okada, T. Kurosaki, Regulated selection of germinal-center cells into the memory B cell compartment. *Nat Immunol*. 17, 861–869 (2016).

34. U. Klein, K. Rajewsky, R. Küppers, Human Immunoglobulin (Ig)M+IgD+ Peripheral Blood B Cells Expressing the CD27 Cell Surface Antigen Carry Somatically Mutated Variable Region Genes: CD27 as a General Marker for Somatically Mutated (Memory) B Cells. *J Exp Medicine*. 188, 1679–1689 (1998).

35. C. T. Mayer, A. Gazumyan, E. E. Kara, A. D. Gitlin, J. Golijanin, C. Viant, J. Pai, T. Y. Oliveira, Q. Wang, A. Escolano, M. Medina-Ramirez, R. W. Sanders, M. C. Nussenzweig, The microanatomic segregation of selection by apoptosis in the germinal center. *Science*. 358 (2017), doi:10.1126/science.aao2602.

36. A. Adhikari, P. Mainali, J. K. Davie, JARID2 and the PRC2 complex regulate the cell cycle in skeletal muscle. *J Biol Chem*. 294, 19451–19464 (2019).

37. I. Sanz, C. Wei, S. A. Jenks, K. S. Cashman, C. Tipton, M. C. Woodruff, J. Hom, F. E.-H. Lee, Challenges and Opportunities for Consistent Classification of Human B Cell and Plasma Cell Populations. *Front Immunol*. 10, 2458 (2019).

38. R. Shinkura, S. Ito, N. A. Begum, H. Nagaoka, M. Muramatsu, K. Kinoshita, Y. Sakakibara, H. Hijikata, T. Honjo, Separate domains of AID are required for somatic hypermutation and class-switch recombination. *Nat Immunol*. 5, 707–712 (2004).

39. J. H. Choi, K. Wang, D. Zhang, X. Zhan, T. Wang, C.-H. Bu, C. L. Behrendt, M. Zeng, Y. Wang, T. Misawa, X. Li, M. Tang, X. Zhan, L. Scott, S. Hildebrand, A. R. Murray, E. M. Y. Moresco, L. V. Hooper, B. Beutler, IgD class switching is initiated by microbiota and limited to mucosa-associated lymphoid tissue in mice. *Proc National Acad Sci*. 114, E1196–E1204 (2017).

40. D. Jung, C. Giallourakis, R. Mostoslavsky, F. W. Alt, MECHANISM AND CONTROL OF V(D)J RECOMBINATION AT THE IMMUNOGLOBULIN HEAVY CHAIN LOCUS. *Annu Rev Immunol*. 24, 541–570 (2006).

41. P. Early, H. Huang, M. Davis, K. Calame, L. Hood, An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. *Cell*. 19, 981–992 (1980).

42. R. M. Brecht, C. C. Liu, H. A. Beilinson, A. Khitun, S. A. Slavoff, D. G. Schatz, Nucleolar localization of RAG1 modulates V(D)J recombination activity. *Proc National Acad Sci*. 117, 4300–4309 (2020).

43. L. Zhang, T. L. Reynolds, X. Shan, S. Desiderio, Coupling of V(D)J Recombination to the Cell Cycle Suppresses Genomic Instability and Lymphoid Tumorigenesis. *Immunity*. 34, 163–174 (2011).

44. T. Gan, Y. Wang, Y. Liu, D. G. Schatz, J. Hu, RAG2 abolishes RAG1 aggregation to facilitate V(D)J recombination. *Cell Reports*. 37, 109824 (2021).

45. F. F. Yin, S. Bailey, C. A. Innis, M. Ciubotaru, S. Kamtekar, T. A. Steitz, D. G. Schatz, Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis. *Nat Struct Mol Biol*. 16, 499–508 (2009).

46. Y. Ma, U. Pannicke, K. Schwarz, M. R. Lieber, Hairpin Opening and Overhang Processing by an Artemis/DNA-Dependent Protein Kinase Complex in Nonhomologous End Joining and V(D)J Recombination. *Cell*. 108, 781–794 (2002).

47. H. Lu, K. Schwarz, M. R. Lieber, Extent to which hairpin opening by the Artemis:DNA-PKcs complex can contribute to junctional diversity in V(D)J recombination. *Nucleic Acids Res*. 35, 6917–6923 (2007).

48. J. A. E. Repasky, E. Corbett, C. Boboila, D. G. Schatz, Mutational Analysis of Terminal Deoxynucleotidyltransferase- Mediated N-Nucleotide Addition in V(D)J Recombination. *J Immunol*. 172, 5478–5488 (2004).

49. A. Libri, T. Marton, L. Deriano, The (Lack of) DNA Double-Strand Break Repair Pathway Choice During V(D)J Recombination. *Frontiers Genetics*. 12, 823943 (2022).

50. P. M. Kirkham, H. W. Schroeder, Antibody structure and the evolution of immunoglobulin V gene segments. *Semin Immunol*. 6, 347–360 (1994).

51. J. L. Xu, M. M. Davis, Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities. *Immunity*. 13, 37–45 (2000).

52. D. Noel, T. Bernardi, I. Navarro-Teulon, M. Marin, J.-P. Martinetto, F. Ducancel, J.-C. Mani, B. Pau, M. Piechaczyk, M. Biard-Piechaczyk, Analysis of the individual contributions of immunoglobulin heavy and light chains to the binding of antigen using cell transfection and plasmon resonance analysis. *J Immunol Methods*. 193, 177–187 (1996).

53. A. K. Stewart, R. S. Schwartz, Immunoglobulin V regions and the B cell. *Blood*. 83, 1717–30 (1994).

54. C. Oudinet, X. Zhang, N. Puget, N. Kyritsis, C. Leduc, F.-Z. Braikia, A. Dauba, F. W. Alt, A. A. Khamlichi, Switch Tandem Repeats Influence the Choice of the Alternative End-Joining Pathway in Immunoglobulin Class Switch Recombination. *Front Immunol*. 13, 870933 (2022).

55. S. Zheng, B. Q. Vuong, B. Vaidyanathan, J.-Y. Lin, F.-T. Huang, J. Chaudhuri, Non-coding RNA Generated following Lariat Debranching Mediates Targeting of AID to DNA. *Cell*. 161, 762–773 (2015).

56. S. G. Tangye, A. Ferguson, D. T. Avery, C. S. Ma, P. D. Hodgkin, Isotype Switching by Human B Cells Is Division-Associated and Regulated by Cytokines. *J Immunol*. 169, 4298–4306 (2002).

57. Y. Kawano, T. Noma, K. Kou, I. Yoshizawa, J. Yata, Regulation of human IgG subclass production by cytokines: human IgG subclass production enhanced differentially by interleukin-6. *Immunology*. 84, 278–84 (1995).

58. G. Vidarsson, G. Dekkers, T. Rispens, IgG Subclasses and Allotypes: From Structure to Effector Functions. *Front Immunol*. 5, 520 (2014).

59. A. Cerutti, M. Rescigno, The Biology of Intestinal Immunoglobulin A Responses. *Immunity*. 28, 740–750 (2008).

60. C. Gutzeit, K. Chen, A. Cerutti, The enigmatic function of IgD: some answers at last. *Eur. J. Immunol.* 48, 1101–1113 (2018).

61. L. Han, S. Masani, K. Yu, Overlapping activation-induced cytidine deaminase hotspot motifs in Ig class-switch recombination. *Proc National Acad Sci*. 108, 11584–11589 (2011).

62. A. L. Kenter, S. Kumar, R. Wuerffel, F. Grigera, AID hits the jackpot when missing the target. *Curr Opin Immunol*. 39, 96–102 (2016).

63. S. P. Methot, J. M. D. Noia, Chapter Two Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. *Adv Immunol*. 133, 37–87 (2017).

64. X. Zhang, Y. Zhang, Z. Ba, N. Kyritsis, R. Casellas, F. W. Alt, Fundamental roles of chromatin loop extrusion in antibody class switching. *Nature*. 575, 385–389 (2019).

65. J. Medvedovic, A. Ebert, H. Tagoh, I. M. Tamir, T. A. Schwickert, M. Novatchkova, Q. Sun, P. J. Huis in 't Veld, C. Guo, H. S. Yoon, Y. Denizot, S. J. B. Holwerda, W. de Laat, M. Cogné, Y. Shi, F. W. Alt, M. Busslinger, Flexible Long-Range Loops in the VH Gene Region of the Igh Locus Facilitate the Generation of a Diverse Antibody Repertoire. *Immunity*. 39, 229–244 (2013).

66. S. Feldman, I. Achour, R. Wuerffel, S. Kumar, T. Gerasimova, R. Sen, A. L. Kenter, Constraints Contributed by Chromatin Looping Limit Recombination Targeting during Ig Class Switch Recombination. *J Immunol*. 194, 2380–2389 (2015).

67. B. K. Birshtein, The role of CTCF binding sites in the 3′ immunoglobulin heavy chain regulatory region. *Frontiers Genetics*. 3, 251 (2012).

68. H. M. Shen, R. Wuerffel, J. F. Cantillo, S. Priyadarshi, X. Lei, J. Liang, Y. L. Wu, A. L. Kenter, Loop extrusion promotes an alternate pathway for isotype switching. *Cell Reports*. 37, 110059 (2021).

69. T. Saha, D. Sundaravinayagam, M. D. Virgilio, Charting a DNA Repair Roadmap for Immunoglobulin Class Switch Recombination. *Trends Biochem Sci*. 46, 184–199 (2020).

70. Z. Xu, H. Zan, E. J. Pone, T. Mai, P. Casali, Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nat Rev Immunol*. 12, 517–531 (2012).

71. U. von Schwedler, H.-M. Jck, M. Wabl, Circular DNA is a product of the immunoglobulin class switch rearrangement. *Nature*. 345, 452–456 (1990).

72. J. A. Roco, L. Mesin, S. C. Binder, C. Nefzger, P. Gonzalez-Figueroa, P. F. Canete, J. Ellyard, Q. Shen, P. A. Robert, J. Cappello, H. Vohra, Y. Zhang, C. R. Nowosad, A. Schiepers, L. M. Corcoran, K.-M. Toellner, J. M. Polo, M. Meyer-Hermann, G. Victora, C. G. Vinuesa, Class-Switch Recombination Occurs Infrequently in Germinal Centers. *Immunity* (2019), doi:10.1016/j.immuni.2019.07.001.

73. H. Xiong, J. Dolpady, M. Wabl, M. A. C. de Lafaille, J. J. Lafaille, Sequential class switching is required for the generation of high affinity IgE antibodies. *J Exp Medicine*. 209, 353–364 (2012).

74. M. C. Zelm, B cells take their time: sequential IgG class switching over the course of an immune response? *Immunol Cell Biol*. 92, 645–646 (2014).

75. M. Lin, L. Du, P. Brandtzaeg, Q. Pan-Hammarström, IgA subclass switch recombination in human mucosal and systemic immune compartments. *Mucosal Immunol*. 7, 511–520 (2014).

76. M. Nakamura, S. Kondo, M. Sugai, M. Nazarea, S. Imamura, T. Honjo, High frequency class switching of an lgM+ B lymphoma clone CH12F3 to lgA+ cells. *Int Immunol*. 8, 193–201 (1996).

77. R. A. Panchakshari, X. Zhang, V. Kumar, Z. Du, P.-C. Wei, J. Kao, J. Dong, F. W. Alt, DNA double-strand break response factors influence end-joining features of IgH class switch and general translocation junctions. *Proc National Acad Sci*. 115, 762–767 (2018).

78. P. P. Rocha, R. Raviram, Y. Fu, J. Kim, V. M. Luo, A. Aljoufi, E. Swanzey, A. Pasquarella, A. Balestrini, E. R. Miraldi, R. Bonneau, J. Petrini, G. Schotta, J. A. Skok, A Damage-Independent Role for 53BP1 that Impacts Break Order and Igh Architecture during Class Switch Recombination. *Cell Reports*. 16, 48–55 (2016).

79. L. Vian, A. Pękowska, S. S. P. Rao, K.-R. Kieffer-Kwon, S. Jung, L. Baranello, S.-C. Huang, L. E. Khattabi, M. Dose, N. Pruett, A. L. Sanborn, A. Canela, Y. Maman, A. Oksanen, W. Resch, X. Li, B. Lee, A. L. Kovalchuk, Z. Tang, S. Nelson, M. D. Pierro, R. R. Cheng, I. Machol, B. G. S. Hilaire, N. C. Durand, M. S. Shamim, E. K. Stamenova, J. N. Onuchic, Y. Ruan, A. Nussenzweig, D. Levens, E. L. Aiden, R. Casellas, The Energetics and Physiological Impact of Cohesin Extrusion. *Cell*. 173, 1165-1178.e20 (2018).

80. C. Tang, D. Bagnara, N. Chiorazzi, M. D. Scharff, T. MacCarthy, AID Overlapping and Polη Hotspots Are Key Features of Evolutionary Variation Within the Human Antibody Heavy Chain (IGHV) Genes. *Front Immunol*. 11, 788 (2020).

81. B. Pilzecker, H. Jacobs, Mutating for Good: DNA Damage Responses During Somatic Hypermutation. *Front Immunol*. 10, 438 (2019).

82. H. Islam, M. Kobayashi, T. Honjo, Apurinic/apyrimidinic endonuclease 1 (APE1) is dispensable for activation-induced cytidine deaminase (AID)-dependent somatic hypermutation in the immunoglobulin gene. *Int Immunol*. 31, 543–554 (2019).

83. F. Delbos, A. D. Smet, A. Faili, S. Aoufouchi, J.-C. Weill, C.-A. Reynaud, Contribution of DNA polymerase η to immunoglobulin gene hypermutation in the mouse. *J Exp Medicine*. 201, 1191–1196 (2005).

84. D. Nemazee, Mechanisms of central tolerance for B cells. *Nat Rev Immunol*. 17, 281–294 (2017).

85. J. S. Rice, J. Newman, C. Wang, D. J. Michael, B. Diamond, Receptor editing in peripheral B cell tolerance. *Proc National Acad Sci*. 102, 1608–1613 (2005).

86. G. V. Borzillo, M. D. Cooper, H. Kubagawa, A. Landay, P. D. Burrows, Isotype switching in human B lymphocyte malignancies occurs by DNA deletion: evidence for nonspecific switch recombination. *J Immunol Baltim Md 1950*. 139, 1326–35 (1987).

87. E. Edry, H. Azulay-Debby, D. Melamed, TOLL-like receptor ligands stimulate aberrant class switch recombination in early B cell precursors. *Int Immunol*. 20, 1575–1585 (2008).

88. L. Delpy, M. L. Bert, M. Cogné, A. A. Khamlichi, Germ-line transcription occurs on both the functional and the non-functional alleles of immunoglobulin constant heavy chain genes. *Eur. J. Immunol.* 33, 2108–2113 (2003).

89. E. Sonoda, Y. Pewzner-Jung, S. Schwers, S. Taki, S. Jung, D. Eilat, K. Rajewsky, B Cell Development under the Condition of Allelic Inclusion. *Immunity*. 6, 225–233 (1997).

90. H. Cedar, Y. Bergman, Choreography of Ig allelic exclusion. *Curr Opin Immunol*. 20, 308–317 (2008).

91. R. Levin-Klein, Y. Bergman, Epigenetic Regulation of Monoallelic Rearrangement (Allelic Exclusion) of Antigen Receptor Genes. *Front Immunol*. 5, 625 (2014).

92. C. Vettermann, M. S. Schlissel, Allelic exclusion of immunoglobulin genes: models and mechanisms. *Immunol Rev*. 237, 22–42 (2010).

93. J. Tan, K. Pieper, L. Piccoli, A. Abdi, M. Foglierini, R. Geiger, C. M. Tully, D. Jarrossay, F. M. Ndungu, J. Wambua, P. Bejon, C. S. Fregni, B. Fernandez-Rodriguez, S. Barbieri, S. Bianchi, K. Marsh, V. Thathy, D. Corti, F. Sallusto, P. Bull, A. Lanzavecchia, A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature*. 529, 105 (2016).

94. K. Pieper, J. Tan, L. Piccoli, M. Foglierini, S. Barbieri, Y. Chen, C. Silacci-Fregni, T. Wolf, D. Jarrossay, M. Anderle, A. Abdi, F. M. Ndungu, O. K. Doumbo, B. Traore, T. M. Tran, S. Jongo, I. Zenklusen, P. D. Crompton, C. Daubenberger, P. C. Bull, F. Sallusto, A. Lanzavecchia, Public antibodies to malaria antigens generated by two LAIR1 insertion modalities. *Nature*. 548, 597 (2017).

95. Y. Chen, K. Xu, L. Piccoli, M. Foglierini, J. Tan, W. Jin, J. Gorman, Y. Tsybovsky, B. Zhang, B. Traore, C. Silacci-Fregni, C. Daubenberger, P. D. Crompton, R. Geiger, F. Sallusto, P. D. Kwong, A. Lanzavecchia, Structural basis of malaria RIFIN binding by LILRB1-containing antibodies. *Nature*. 592, 639–643 (2021).

96. K. Xu, Y. Wang, C.-H. Shen, Y. Chen, B. Zhang, K. Liu, Y. Tsybovsky, S. Wang, S. K. Farney, J. Gorman, T. Stephens, R. Verardi, Y. Yang, T. Zhou, G.-Y. Chuang, A. Lanzavecchia, L. Piccoli, P. D. Kwong, Structural basis of LAIR1 targeting by polymorphic Plasmodium RIFINs. *Nat Commun*. 12, 4226 (2021).

97. M. Lebedin, M. Foglierini, S. Khorkova, C. V. García, C. Ratswohl, A. N. Davydov, M. A. Turchaninova, C. Daubenberger, D. M. Chudakov, A. Lanzavecchia, K. de la Rosa, Different classes of genomic inserts contribute to human antibody diversity. *Proc National Acad Sci*. 119 (2022), doi:10.1073/pnas.2205470119.

98. M. van den Bosch, R. T. Bree, N. F. Lowndes, The MRN complex: coordinating and mediating the response to broken chromosomes. *Embo Rep*. 4, 844–849 (2003).

99. L. Bian, Y. Meng, M. Zhang, D. Li, MRE11-RAD50-NBS1 complex alterations and DNA damage response: implications for cancer treatment. *Mol Cancer*. 18, 169 (2019).

100. A. Kinner, W. Wu, C. Staudt, G. Iliakis, γ-H2AX in recognition and signaling of DNA double-strand breaks in the context of chromatin. *Nucleic Acids Res*. 36, 5678–5694 (2008).

101. Z. Mao, M. Bozzella, A. Seluanov, V. Gorbunova, DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle*. 7, 2902–2906 (2008).

102. Z. Mao, M. Bozzella, A. Seluanov, V. Gorbunova, Comparison of nonhomologous end joining and homologous recombination in human cells. *Dna Repair*. 7, 1765–1771 (2008).

103. A. Bothmer, D. F. Robbiani, M. Di Virgilio, S. F. Bunting, I. A. Klein, N. Feldhahn, J. Barlow, H.-T. Chen, D. Bosque, E. Callen, A. Nussenzweig, M. C. Nussenzweig, Regulation of DNA End Joining, Resection, and Immunoglobulin Class Switch Recombination by 53BP1. *Mol Cell*. 42, 319–329 (2011).

104. L. Feng, K.-W. Fong, J. Wang, W. Wang, J. Chen, RIF1 Counteracts BRCA1-mediated End Resection during DNA Repair*. *J Biol Chem*. 288, 11135–11143 (2013).

105. M. D. Virgilio, E. Callen, A. Yamane, W. Zhang, M. Jankovic, A. D. Gitlin, N. Feldhahn, W. Resch, T. Y. Oliveira, B. T. Chait, A. Nussenzweig, R. Casellas, D. F. Robbiani, M. C. Nussenzweig, Rif1 Prevents Resection of DNA Breaks and Promotes Immunoglobulin Class Switching. *Science*. 339, 711–715 (2013).

106. M. L. Swift, K. Beishline, J. Azizkhan-Clifford, Sp1-dependent recruitment of the histone acetylase p300 to DSBs facilitates chromatin remodeling and recruitment of the NHEJ repair factor Ku70. *Dna Repair*. 105, 103171 (2021).

107. E. Weterings, D. J. Chen, The endless tale of non-homologous end-joining. *Cell Res*. 18, 114–124 (2008).

108. B. J. Davis, J. M. Havener, D. A. Ramsden, End-bridging is required for pol μ to efficiently promote repair of noncomplementary ends by nonhomologous end joining. *Nucleic Acids Res*. 36, 3085–3094 (2008).

109. Y. Wang, B. J. Lamarche, M.-D. Tsai, Human DNA Ligase IV and the Ligase IV/XRCC4 Complex: Analysis of Nick Ligation Fidelity †. *Biochemistry-us*. 46, 4962–4976 (2007).

110. X. Xiong, Z. Du, Y. Wang, Z. Feng, P. Fan, C. Yan, H. Willers, J. Zhang, 53BP1 promotes microhomology-mediated end-joining in G1-phase cells. *Nucleic Acids Res*. 43, 1659–1670 (2015).

111. A. Sallmyr, A. E. Tomkinson, Repair of DNA double-strand breaks by mammalian alternative end-joining pathways. *J Biol Chem*. 293, 10536–10546 (2018).

112. A. Sfeir, L. S. Symington, Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem Sci*. 40, 701–714 (2015).

113. H. Zan, C. Tat, Z. Qiu, J. R. Taylor, J. A. Guerrero, T. Shen, P. Casali, Rad52 competes with Ku70/Ku86 for binding to S-region DSB ends to modulate antibody class-switch DNA recombination. *Nat Commun*. 8, 14244 (2017).

114. D. Simsek, M. Jasin, Alternative end-joining is suppressed by the canonical NHEJ component Xrcc4–ligase IV during chromosomal translocation formation. *Nat Struct Mol Biol*. 17, 410–416 (2010).

115. M. McVey, S. E. Lee, MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet*. 24, 529–538 (2008).

116. S.-W. Yoon, D.-K. Kim, K. P. Kim, K.-S. Park, Rad51 Regulates Cell Cycle Progression by Preserving G2/M Transition in Mouse Embryonic Stem Cells. *Stem Cells Dev*. 23, 2700–2711 (2014).

117. I. E. Wassing, E. Graham, X. Saayman, L. Rampazzo, C. Ralf, A. Bassett, F. Esashi, The RAD51 recombinase protects mitotic chromatin in human cells. *Nat Commun*. 12, 5380 (2021).

118. A. So, E. Dardillac, A. Muhammad, C. Chailleux, L. Sesma-Sanz, S. Ragu, E. Le Cam, Y. Canitrot, J. Y. Masson, P. Dupaigne, B. S. Lopez, J. Guirouilh-Barbat, RAD51 protects against nonconservative DNA double-strand break repair through a nonenzymatic function. *Nucleic Acids Res*. 50, gkac073- (2022).

119. R. Scully, A. Panday, R. Elango, N. A. Willis, DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat Rev Mol Cell Bio*. 20, 698–714 (2019).

120. M. A. Jafri, S. A. Ansari, M. H. Alqahtani, J. W. Shay, Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies. *Genome Med*. 8, 69 (2016).

121. E. Bonnell, E. Pasquier, R. J. Wellinger, Telomere Replication: Solving Multiple End Replication Problems. *Frontiers Cell Dev Biology*. 9, 668171 (2021).

122. B. T. Hu, S. C. Lee, E. Marin, D. H. Ryan, R. A. Insel, Telomerase is up-regulated in human germinal center B cells in vivo and can be re-expressed in memory B cells activated in vitro. *J Immunol Baltim Md 1950*. 159, 1068–71 (1997).

123. N. Weng, L. Granger, R. J. Hodes, Telomere lengthening and telomerase activation during human B cell differentiation. *Proc National Acad Sci*. 94, 10827–10832 (1997).

124. C. M. Roake, S. E. Artandi, Regulation of human telomerase in homeostasis and disease. *Nat Rev Mol Cell Bio*. 21, 384–397 (2020).

125. K. Okamoto, H. Seimiya, Revisiting Telomere Shortening in Cancer. *Cells*. 8, 107 (2019).

126. E. Fouquerel, D. Parikh, P. Opresko, DNA damage processing at telomeres: The ends justify the means. *Dna Repair*. 44, 159–168 (2016).

127. J. W. Parks, M. D. Stone, Coordinated DNA dynamics during the human telomerase catalytic cycle. *Nat Commun*. 5, 4146 (2014).

128. T. Aramburu, S. Plucinsky, E. Skordalakes, Pot1-tpp1 telomere length regulation and disease. *Comput Struct Biotechnology J*. 18, 1939–1946 (2020).

129. S. G. Nergadze, M. A. Santagostino, A. Salzano, C. Mondello, E. Giulotto, Contribution of telomerase RNA retrotranscription to DNA double-strand break repair during mammalian genome evolution. *Genome Biol*. 8, R260 (2007).

130. L. Sola, S. G. Nergadze, E. Cappelletti, F. M. Piras, E. Giulotto, M. Santagostino, Telomeric-Like Repeats Flanked by Sequences Retrotranscribed from the Telomerase RNA Inserted at DNA Double-Strand Break Sites during Vertebrate Genome Evolution. *Int J Mol Sci*. 22, 11048 (2021).

131. J. H. Santos, J. N. Meyer, M. Skorvaga, L. A. Annab, B. V. Houten, Mitochondrial hTERT exacerbates free-radical-mediated mtDNA damage. *Aging Cell*. 3, 399–411 (2004).

132. J. Rosen, P. Jakobs, N. Ale-Agha, J. Altschmied, J. Haendeler, Non-canonical functions of Telomerase Reverse Transcriptase – Impact on redox homeostasis. *Redox Biol*. 34, 101543 (2020).

133. K. Kiwerska, K. Szyfter, DNA repair in cancer initiation, progression, and therapy—a double-edged sword. *J Appl Genetics*. 60, 329–334 (2019).

134. J. C. Roach, G. Glusman, A. F. A. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, D. J. Galas, Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science*. 328, 636–639 (2010).

135. M. O'Driscoll, K. M. Cerosaletti, P.-M. Girard, Y. Dai, M. Stumm, B. Kysela, B. Hirsch, A. Gennery, S. E. Palmer, J. Seidel, R. A. Gatti, R. Varon, M. A. Oettinger, H. Neitzel, P. A. Jeggo, P. Concannon, DNA Ligase IV Mutations Identified in Patients Exhibiting Developmental Delay and Immunodeficiency. *Mol Cell*. 8, 1175–1185 (2001).

136. Q. Pan-Hammarström, A.-M. Jones, A. Lähdesmäki, W. Zhou, R. A. Gatti, L. Hammarström, A. R. Gennery, M. R. Ehrenstein, Impact of DNA ligase IV on nonhomologous end joining pathways during class switch recombination in human cells. *J Exp Medicine*. 201, 189–194 (2005).

137. C. D. McFarland, J. A. Yaglom, J. W. Wojtkowiak, J. G. Scott, D. L. Morse, M. Y. Sherman, L. A. Mirny, The Damaging Effect of Passenger Mutations on Cancer Progression. *Cancer Res*. 77, 4763–4772 (2017).

138. M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, G. Getz, Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 505, 495–501 (2014).

139. B. A. Baldo, N. H. Pham, Adverse reactions to targeted and non-targeted chemotherapeutic drugs with emphasis on hypersensitivity responses and the invasive metastatic switch. *Cancer Metast Rev*. 32, 723–761 (2013).

140. E. Kuijk, O. Kranenburg, E. Cuppen, A. V. Hoeck, Common anti-cancer therapies induce somatic mutations in stem cells of healthy tissue. *Nat Commun*. 13, 5915 (2022).

141. M. A. Sherman, A. U. Yaari, O. Priebe, F. Dietlein, P.-R. Loh, B. Berger, Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nat Biotechnol*, 1–10 (2022).

142. T. T. Wu, M. Reid-Miller, H. M. Perry, E. A. Kabat, Long identical repeats in the mouse gamma 2b switch region and their implications for the mechanism of class switching. *Embo J*. 3, 2033–2040 (1984).

143. K. A. Lukyanov, G. A. Launer, V. S. Tarabykin, A. G. Zaraisky, S. A. Lukyanov, Inverted Terminal Repeats Permit the Average Length of Amplified DNA Fragments to Be Regulated during Preparation of cDNA Libraries by Polymerase Chain Reaction. *Anal Biochem*. 229, 198–202 (1995).

144. F. Boyer, H. Boutouil, I. Dalloul, Z. Dalloul, J. Cook-Moreau, J.-C. Aldigier, C. Carrion, B. Herve, E. Scaon, M. Cogné, S. Péron, CSReport: A New Computational Tool Designed for Automatic Analysis of Class Switch Recombination Junctions Sequenced by High-Throughput Sequencing. *J Immunol*. 198, 4148–4155 (2017).

145. H. Jia, Y. Guo, W. Zhao, K. Wang, Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. *Sci Rep-uk*. 4, 5737 (2014).

146. J. L. Weirather, M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X.-J. Wang, D. Buck, K. F. Au, Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000research*. 6, 100 (2017).

147. Z. Udaondo, K. Sittikankaew, T. Uengwetwanit, T. Wongsurawat, C. Sonthirod, P. Jenjaroenpun, W. Pootakham, N. Karoonuthaisiri, I. Nookaew, Comparative Analysis of PacBio and Oxford Nanopore Sequencing Technologies for Transcriptomic Landscape Identification of Penaeus monodon. *Life*. 11, 862 (2021).

148. J. Cui, N. shen, Z. Lu, G. Xu, Y. Wang, B. Jin, Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. *Plant Methods*. 16, 85 (2020).

149. P. M. Ashton, S. Nair, T. Dallman, S. Rubino, W. Rabsch, S. Mwaigwisya, J. Wain, J. O'Grady, MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol*. 33, 296–300 (2015).

150. M. Philpott, J. Watson, A. Thakurta, T. Brown, T. Brown, U. Oppermann, A. P. Cribbs, Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. *Nat Biotechnol*, 1–4 (2021).

151. C. Delahaye, J. Nicolas, Sequencing DNA with nanopores: Troubles and biases. *Plos One*. 16, e0257521 (2021).

152. M. Jain, H. E. Olsen, B. Paten, M. Akeson, The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*. 17, 239 (2016).

153. J. Stavnezer, J. E. J. Guikema, C. E. Schrader, Mechanism and Regulation of Class Switch Recombination. *Annu Rev Immunol*. 26, 261–292 (2008).

154. R. J. M. Bashford-Rogers, L. Bergamaschi, E. F. McKinney, D. C. Pombal, F. Mescia, J. C. Lee, D. C. Thomas, S. M. Flint, P. Kellam, D. R. W. Jayne, P. A. Lyons, K. G. C. Smith, Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature*. 574, 122–126 (2019).

155. M. A. Turchaninova, A. Davydov, O. V. Britanova, M. Shugay, V. Bikos, E. S. Egorov, V. I. Kirgizova, E. M. Merzlyak, D. B. Staroverov, D. A. Bolotin, I. Z. Mamedov, M. Izraelson, M. D. Logacheva, O. Kladova, K. Plevova, S. Pospisilova, D. M. Chudakov, High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc*. 11, 1599–1616 (2016).

156. O. Hammarsten, A. Lyytikäinen, S. Thunström, T. Ek, A. Fasth, O. Ekwall, S. Cajander, E. W. Borgström, C. I. E. Smith, P. Johansson, Clinical measurement of cellular DNA damage hypersensitivity in patients with DNA repair defects. *Orphanet J Rare Dis*. 17, 50 (2022).

157. S. Bhattacharjee, S. Nandi, Rare Genetic Diseases with Defects in DNA Repair: Opportunities and Challenges in Orphan Drug Development for Targeted Cancer Therapy. *Cancers*. 10, 298 (2018).

158. D. Sundaravinayagam, A. Rahjouei, M. Andreani, D. Tupiņa, S. Balasubramanian, T. Saha, V. Delgado-Benito, V. Coralluzzo, O. Daumke, M. D.

Virgilio, 53BP1 Supports Immunoglobulin Class Switch Recombination Independently of Its DNA Double-Strand Break End Protection Function. *Cell Reports*. 28, 1389-1399.e6 (2019).

159. V. Delgado-Benito, D. B. Rosen, Q. Wang, A. Gazumyan, J. A. Pai, T. Y. Oliveira, D. Sundaravinayagam, W. Zhang, M. Andreani, L. Keller, K.-R. Kieffer-Kwon, A. Pękowska, S. Jung, M. Driesner, R. I. Subbotin, R. Casellas, B. T. Chait, M. C. Nussenzweig, M. D. Virgilio, The Chromatin Reader ZMYND8 Regulates Igh Enhancers to Promote Immunoglobulin Class Switch Recombination. *Mol Cell*. 72 (2018), doi:10.1016/j.molcel.2018.08.042.

160. A. Reboldi, T. I. Arnon, L. B. Rodda, A. Atakilit, D. Sheppard, J. G. Cyster, IgA production requires B cell interaction with subepithelial dendritic cells in Peyer's patches. *Science*. 352, aaf4822 (2016).

161. S. Hamperl, M. J. Bocek, J. C. Saldivar, T. Swigut, K. A. Cimprich, Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell*. 170, 774-786.e19 (2017).

162. J. H. Barlow, R. B. Faryabi, E. Callén, N. Wong, A. Malhowski, H. T. Chen, G. Gutierrez-Cruz, H.-W. Sun, P. McKinnon, G. Wright, R. Casellas, D. F. Robbiani, L. Staudt, O. Fernandez-Capetillo, A. Nussenzweig, Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. *Cell*. 152, 620–632 (2013).

163. Z. Hu, Z. Shi, X. Guo, B. Jiang, G. Wang, D. Luo, Y. Chen, Y.-S. Zhu, Ligase IV inhibitor SCR7 enhances gene editing directed by CRISPR–Cas9 and ssODN in human cancer cells. *Cell Biosci*. 8, 12 (2018).

164. R. Bhargava, D. O. Onyango, J. M. Stark, Regulation of Single-Strand Annealing and its Role in Genome Maintenance. *Trends Genet*. 32, 566–575 (2016).

165. S. Matsuoka, B. A. Ballif, A. Smogorzewska, E. R. M. III, K. E. Hurov, J. Luo, C. E. Bakalarski, Z. Zhao, N. Solimini, Y. Lerenthal, Y. Shiloh, S. P. Gygi, S. J. Elledge, ATM and ATR Substrate Analysis Reveals Extensive Protein Networks Responsive to DNA Damage. *Science*. 316, 1160–1166 (2007).

166. G. B. Morin, The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. *Cell*. 59, 521–529 (1989).

167. E. Pascolo, C. Wenz, J. Lingner, N. Hauel, H. Priepke, I. Kauffmann, P. Garin-Chesa, W. J. Rettig, K. Damm, A. Schnapp, Mechanism of Human Telomerase Inhibition by BIBR1532, a Synthetic, Non-nucleosidic Drug Candidate*. *J Biol Chem*. 277, 15566–15572 (2002).

168. H. El-Daly, M. Kull, S. Zimmermann, M. Pantic, C. F. Waller, U. M. Martens, Selective cytotoxicity and telomere damage in leukemia cells using the telomerase inhibitor BIBR1532. *Blood*. 105, 1742–1749 (2005).

169. G. Altamura, B. degli Uberti, G. Galiero, G. D. Luca, K. Power, L. Licenziato, P. Maiolino, G. Borzacchiello, The Small Molecule BIBR1532 Exerts Potential Anti-cancer Activities in Preclinical Models of Feline Oral Squamous Cell Carcinoma Through Inhibition of Telomerase Activity and Down-Regulation of TERT. *Frontiers Vet Sci*. 7, 620776 (2021).

170. D. Menolfi, S. Zha, ATM, ATR and DNA-PKcs kinases—the lessons from the mouse models: inhibition ≠ deletion. *Cell Biosci*. 10, 8 (2020).

171. K. L. Sheaffer, E. N. Elliott, K. H. Kaestner, DNA Hypomethylation Contributes to Genomic Instability and Intestinal Cancer Initiation. *Cancer Prev Res*. 9, 534–546 (2016).

172. M. Kulis, A. Merkel, S. Heath, A. C. Queirós, R. P. Schuyler, G. Castellano, R. Beekman, E. Raineri, A. Esteve, G. Clot, N. Verdaguer-Dot, M. Duran-Ferrer, N. Russiñol, R. Vilarrasa-Blasi, S. Ecker, V. Pancaldi, D. Rico, L. Agueda, J. Blanc, D. Richardson, L. Clarke, A. Datta, M. Pascual, X. Agirre, F. Prosper, D. Alignani, B. Paiva, G. Caron, T. Fest, M. O. Muench, M. E. Fomin, S.-T. Lee, J. L. Wiemels, A. Valencia, M. Gut, P. Flicek, H. G. Stunnenberg, R. Siebert, R. Küppers, I. G. Gut, E. Campo, J. I. Martín-Subero, Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat Genet*. 47, 746–756 (2015).

173. L. A. Sanz, S. R. Hartono, Y. W. Lim, S. Steyaert, A. Rajpurkar, P. A. Ginno, X. Xu, F. Chédin, Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol Cell*. 63, 167–178 (2016).

174. B. P. Belotserkovskii, S. Tornaletti, A. D. D'Souza, P. C. Hanawalt, R-loop generation during transcription: formation, processing and cellular outcomes. *Dna Repair* (2018), doi:10.1016/j.dnarep.2018.08.009.

175. M. Špírek, J. Mlčoušková, O. Beláň, M. Gyimesi, G. M. Harami, E. Molnár, J. Novacek, M. Kovács, L. Krejci, Human RAD51 rapidly forms intrinsically dynamic nucleoprotein filaments modulated by nucleotide binding state. *Nucleic Acids Res*. 46, gky111- (2018).

176. E. Enervald, L. Du, T. Visnes, A. Björkman, E. Lindgren, J. Wincent, G. Borck, L. Colleaux, V. Cormier-Daire, D. C. van Gent, J. Pie, B. Puisac, N. F. de Miranda, S. Kracker, L. Hammarström, J.-P. de Villartay, A. Durandy, J. Schoumans, L. Ström, Q. Pan-Hammarström, A regulatory role for the cohesin loader NIPBL in nonhomologous end joining during immunoglobulin class switch recombination. *J Exp Medicine*. 210, 2503–2513 (2013).

177. T. Sun, C. Chen, Y. Wu, S. Zhang, J. Cui, P. Shen, Modeling the role of p53 pulses in DNA damage- induced cell death decision. *Bmc Bioinformatics*. 10, 190 (2009).

178. J. Dong, R. A. Panchakshari, T. Zhang, Y. Zhang, J. Hu, S. A. Volpi, R. M. Meyers, Y.-J. Ho, Z. Du, D. F. Robbiani, F. Meng, M. Gostissa, M. C.

Nussenzweig, J. P. Manis, F. W. Alt, Orientation-specific joining of AID-initiated DNA breaks promotes antibody class switching. *Nature*. 525, 134–139 (2015).

179. X. S. Wang, J. Zhao, F. Wu-Baer, Z. Shao, B. J. Lee, O. M. Cupo, R. Rabadan, J. Gautier, R. Baer, S. Zha, CtIP-mediated DNA resection is dispensable for IgH class switch recombination by alternative end-joining. *Proc National Acad Sci*. 117, 25700–25711 (2020).

180. Y.-J. Kang, C. T. Yan, Regulation of DNA repair in the absence of classical non-homologous end joining. *Dna Repair*. 68, 34–40 (2018).

181. C. Mendez-Dorantes, R. Bhargava, J. M. Stark, Repeat-mediated deletions can be induced by a chromosomal break far from a repeat, but multiple pathways suppress such rearrangements. *Gene Dev*. 32, 524–536 (2018).

182. A. Trenner, A. A. Sartori, Harnessing DNA Double-Strand Break Repair for Cancer Treatment. *Frontiers Oncol*. 9, 1388 (2019).

183. B. van de Kooij, A. Kruswick, H. van Attikum, M. B. Yaffe, Multi-pathway DNA-repair reporters reveal competition between end-joining, single-strand annealing and homologous recombination at Cas9-induced DNA double-strand breaks. *Nat Commun*. 13, 5295 (2022).

184. T. A. Knijnenburg, L. Wang, M. T. Zimmermann, N. Chambwe, G. F. Gao, A. D. Cherniack, H. Fan, H. Shen, G. P. Way, C. S. Greene, Y. Liu, R. Akbani, B. Feng, L. A. Donehower, C. Miller, Y. Shen, M. Karimi, H. Chen, P. Kim, P. Jia, E. Shinbrot, S. Zhang, J. Liu, H. Hu, M. H. Bailey, C. Yau, D. Wolf, Z. Zhao, J. N. Weinstein, L. Li, L. Ding, G. B. Mills, P. W. Laird, D. A. Wheeler, I. Shmulevich, T. C. G. A. R. Network, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson, C. M. Hutter, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. (Julia) Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, S. Meier, M. S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, V. Thorsson, W. Zhang, R. Akbani, B. M. Broom, A. M. Hegde, Z. Ju, R. S. Kanchi, A. Korkut, J. Li, H. Liang, S. Ling, W. Liu, Y. Lu, G. B. Mills, K.-S. Ng, A. Rao, M. Ryan, J. Wang, J. N. Weinstein, J. Zhang, A. Abeshouse, J. Armenia, D. Chakravarty, W. K. Chatila, I. de Bruijn, J. Gao, B. E. Gross, Z. J. Heins, R. Kundra, K. La, M. Ladanyi, A. Luna, M. G. Nissan, A. Ochoa, S. M. Phillips, E. Reznik, F. Sanchez-Vega, C. Sander, N. Schultz, R. Sheridan, S. O. Sumer, Y. Sun, B. S. Taylor, J. Wang, H. Zhang, P. Anur, M. Peto, P. Spellman, C. Benz, J. M. Stuart, C. K. Wong, C. Yau, D. N. Hayes, J. S. Parker, M. D. Wilkerson, A. Ally, M. Balasundaram, R. Bowlby, D. Brooks, R. Carlsen, E. Chuah, N. Dhalla, R. Holt, S. J. M. Jones, K. Kasaian, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, K. Mungall, A. G. Robertson, S. Sadeghi, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, K. Tse, T. Wong, A. C. Berger, R. Beroukhim, A. D. Cherniack, C. Cibulskis, S. B. Gabriel, G. F. Gao, G. Ha, M. Meyerson, S. E. Schumacher, J. Shih, M. H. Kucherlapati, R. S. Kucherlapati, S. Baylin, L. Cope, L. Danilova, M. S. Bootwalla, P. H. Lai, D. T. Maglinte, D. J. V. D. Berg, D. J. Weisenberger, J. T. Auman, S. Balu, T. Bodenheimer, C. Fan, K. A. Hoadley,

A. P. Hoyle, S. R. Jefferys, C. D. Jones, S. Meng, P. A. Mieczkowski, L. E. Mose, A. H. Perou, C. M. Perou, J. Roach, Y. Shi, J. V. Simons, T. Skelly, M. G. Soloway, D. Tan, U. Veluvolu, H. Fan, T. Hinoue, P. W. Laird, H. Shen, W. Zhou, M. Bellair, K. Chang, K. Covington, C. J. Creighton, H. Dinh, H. Doddapaneni, L. A. Donehower, J. Drummond, R. A. Gibbs, R. Glenn, W. Hale, Y. Han, J. Hu, V. Korchina, S. Lee, L. Lewis, W. Li, X. Liu, M. Morgan, D. Morton, D. Muzny, J. Santibanez, M. Sheth, E. Shinbrot, L. Wang, M. Wang, D. A. Wheeler, L. Xi, F. Zhao, J. Hess, E. L. Appelbaum, M. Bailey, M. G. Cordes, L. Ding, C. C. Fronick, L. A. Fulton, R. S. Fulton, C. Kandoth, E. R. Mardis, M. D. McLellan, C. A. Miller, H. K. Schmidt, R. K. Wilson, D. Crain, E. Curley, J. Gardner, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, M. Sherman, E. Thompson, P. Yena, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, N. Corcoran, T. Costello, C. Hovens, A. L. Carvalho, A. C. de Carvalho, J. H. Fregnani, A. Longatto-Filho, R. M. Reis, C. Scapulatempo-Neto, H. C. S. Silveira, D. O. Vidal, A. Burnette, J. Eschbacher, B. Hermes, A. Noss, R. Singh, M. L. Anderson, P. D. Castro, M. Ittmann, D. Huntsman, B. Kohl, X. Le, R. Thorp, C. Andry, E. R. Duffy, V. Lyadov, O. Paklina, G. Setdikova, A. Shabunin, M. Tavobilov, C. McPherson, R. Warnick, R. Berkowitz, D. Cramer, C. Feltmate, N. Horowitz, A. Kibel, M. Muto, C. P. Raut, A. Malykh, J. S. Barnholtz-Sloan, W. Barrett, K. Devine, J. Fulop, Q. T. Ostrom, K. Shimmel, Y. Wolinsky, A. E. Sloan, A. D. Rose, F. Giuliante, M. Goodman, B. Y. Karlan, C. H. Hagedorn, J. Eckman, J. Harr, J. Myers, K. Tucker, L. A. Zach, B. Deyarmin, H. Hu, L. Kvecher, C. Larson, R. J. Mural, S. Somiari, A. Vicha, T. Zelinka, J. Bennett, M. Iacocca, B. Rabeno, P. Swanson, M. Latour, L. Lacombe, B. Têtu, A. Bergeron, M. McGraw, S. M. Staugaitis, J. Chabot, H. Hibshoosh, A. Sepulveda, T. Su, T. Wang, O. Potapova, O. Voronina, L. Desjardins, O. Mariani, S. Roman-Roman, X. Sastre, M.-H. Stern, F. Cheng, S. Signoretti, A. Berchuck, D. Bigner, E. Lipp, J. Marks, S. McCall, R. McLendon, A. Secord, A. Sharp, M. Behera, D. J. Brat, A. Chen, K. Delman, S. Force, F. Khuri, K. Magliocca, S. Maithel, J. J. Olson, T. Owonikoko, A. Pickens, S. Ramalingam, D. M. Shin, G. Sica, E. G. V. Meir, H. Zhang, W. Eijckenboom, A. Gillis, E. Korpershoek, L. Looijenga, W. Oosterhuis, H. Stoop, K. E. van Kessel, E. C. Zwarthoff, C. Calatozzolo, L. Cuppini, S. Cuzzubbo, F. DiMeco, G. Finocchiaro, L. Mattei, A. Perin, B. Pollo, C. Chen, J. Houck, P. Lohavanichbutr, A. Hartmann, C. Stoehr, R. Stoehr, H. Taubert, S. Wach, B. Wullich, W. Kycler, D. Murawa, M. Wiznerowicz, K. Chung, W. J. Edenfield, J. Martin, E. Baudin, G. Bubley, R. Bueno, A. D. Rienzo, W. G. Richards, S. Kalkanis, T. Mikkelsen, H. Noushmehr, L. Scarpace, N. Girard, M. Aymerich, E. Campo, E. Giné, A. L. Guillermo, N. V. Bang, P. T. Hanh, B. D. Phu, Y. Tang, H. Colman, K. Evason, P. R. Dottino, J. A. Martignetti, H. Gabra, H. Juhl, T. Akeredolu, S. Stepa, D. Hoon, K. Ahn, K. J. Kang, F. Beuschlein, A. Breggia, M. Birrer, D. Bell, M. Borad, A. H. Bryce, E. Castle, V. Chandan, J. Cheville, J. A. Copland, M. Farnell, T. Flotte, N. Giama, T. Ho, M. Kendrick, J.-P. Kocher, K. Kopp, C. Moser, D. Nagorney, D. O'Brien, B. P. O'Neill, T. Patel, G. Petersen, F. Que, M. Rivera, L. Roberts, R. Smallridge, T. Smyrk, M. Stanton, R. H. Thompson, M. Torbenson, J. D. Yang, L. Zhang, F. Brimo, J. A. Ajani, A. M. A. Gonzalez, C. Behrens, J. Bondaruk, R. Broaddus, B. Czerniak, B. Esmaeli, J. Fujimoto, J. Gershenwald, C. Guo, A. J. Lazar, C. Logothetis, F. Meric-Bernstam, C. Moran, L. Ramondetta, D. Rice, A. Sood, P. Tamboli, T. Thompson, P. Troncoso, A. Tsao, I. Wistuba, C. Carter, L. Haydu, P. Hersey, V. Jakrot, H. Kakavand, R. Kefford, K. Lee, G. Long, G. Mann, M. Quinn, R. Saw, R. Scolyer, K. Shannon, A. Spillane, J. Stretch, M. Synott, J. Thompson,

J. Wilmott, H. Al-Ahmadie, T. A. Chan, R. Ghossein, A. Gopalan, D. A. Levine, V. Reuter, S. Singer, B. Singh, N. V. Tien, T. Broudy, C. Mirsaidi, P. Nair, P. Drwiega, J. Miller, J. Smith, H. Zaren, J.-W. Park, N. P. Hung, E. Kebebew, W. M. Linehan, A. R. Metwalli, K. Pacak, P. A. Pinto, M. Schiffman, L. S. Schmidt, C. D. Vocke, N. Wentzensen, R. Worrell, H. Yang, M. Moncrieff, C. Goparaju, J. Melamed, H. Pass, N. Botnariuc, I. Caraman, M. Cernat, I. Chemencedji, A. Clipca, S. Doruc, G. Gorincioi, S. Mura, M. Pirtac, I. Stancul, D. Tcaciuc, M. Albert, I. Alexopoulou, A. Arnaout, J. Bartlett, J. Engel, S. Gilbert, J. Parfitt, H. Sekhon, G. Thomas, D. M. Rassl, R. C. Rintoul, C. Bifulco, R. Tamakawa, W. Urba, N. Hayward, H. Timmers, A. Antenucci, F. Facciolo, G. Grazi, M. Marino, R. Merola, R. de Krijger, A.-P. Gimenez-Roqueplo, A. Piché, S. Chevalier, G. McKercher, K. Birsoy, G. Barnett, C. Brewer, C. Farver, T. Naska, N. A. Pennell, D. Raymond, C. Schilero, K. Smolenski, F. Williams, C. Morrison, J. A. Borgia, M. J. Liptay, M. Pool, C. W. Seder, K. Junker, L. Omberg, M. Dinkin, G. Manikhas, D. Alvaro, M. C. Bragazzi, V. Cardinale, G. Carpino, E. Gaudio, D. Chesla, S. Cottingham, M. Dubina, F. Moiseenko, R. Dhanasekaran, K.-F. Becker, K.-P. Janssen, J. Slotta-Huspenina, M. H. Abdel-Rahman, D. Aziz, S. Bell, C. M. Cebulla, A. Davis, R. Duell, J. B. Elder, J. Hilty, B. Kumar, J. Lang, N. L. Lehman, R. Mandt, P. Nguyen, R. Pilarski, K. Rai, L. Schoenfield, K. Senecal, P. Wakely, P. Hansen, R. Lechan, J. Powers, A. Tischler, W. E. Grizzle, K. C. Sexton, A. Kastl, J. Henderson, S. Porten, J. Waldmann, M. Fassnacht, S. L. Asa, D. Schadendorf, M. Couce, M. Graefen, H. Huland, G. Sauter, T. Schlomm, R. Simon, P. Tennstedt, O. Olabode, M. Nelson, O. Bathe, P. R. Carroll, J. M. Chan, P. Disaia, P. Glenn, R. K. Kelley, C. N. Landen, J. Phillips, M. Prados, J. Simko, K. Smith-McCune, S. VandenBerg, K. Roggin, A. Fehrenbach, A. Kendler, S. Sifri, R. Steele, A. Jimeno, F. Carey, I. Forgie, M. Mannelli, M. Carney, B. Hernandez, B. Campos, C. Herold-Mende, C. Jungk, A. Unterberg, A. von Deimling, A. Bossler, J. Galbraith, L. Jacobus, M. Knudson, T. Knutson, D. Ma, M. Milhem, R. Sigmund, A. K. Godwin, R. Madan, H. G. Rosenthal, C. Adebamowo, S. N. Adebamowo, A. Boussioutas, D. Beer, T. Giordano, A.-M. Mes-Masson, F. Saad, T. Bocklage, L. Landrum, R. Mannel, K. Moore, K. Moxley, R. Postier, J. Walker, R. Zuna, M. Feldman, F. Valdivieso, R. Dhir, J. Luketich, E. M. M. Pinero, M. Quintero-Aguilo, C. G. Carlotti, J. S. D. Santos, R. Kemp, A. Sankarankuty, D. Tirapelli, J. Catto, K. Agnew, E. Swisher, J. Creaney, B. Robinson, C. S. Shelley, E. M. Godwin, S. Kendall, C. Shipman, C. Bradford, T. Carey, A. Haddad, J. Moyer, L. Peterson, M. Prince, L. Rozek, G. Wolf, R. Bowman, K. M. Fong, I. Yang, R. Korst, W. K. Rathmell, J. L. Fantacone-Campbell, J. A. Hooke, A. J. Kovatich, C. D. Shriver, J. DiPersio, B. Drake, R. Govindan, S. Heath, T. Ley, B. V. Tine, P. Westervelt, M. A. Rubin, J. I. Lee, N. D. Aredes, A. Mariamidze, R. J. Monnat, Y. Xiao, C. Wang, Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Reports*. 23, 239-254.e6 (2018).

185. J. Ge, L. P. Ngo, S. Kaushal, I. J. Tay, E. Thadhani, J. E. Kay, P. Mazzucato, D. N. Chow, J. L. Fessler, D. M. Weingeist, R. W. Sobol, L. D. Samson, S. R. Floyd, B. P. Engelward, CometChip enables parallel analysis of multiple DNA repair activities. *Dna Repair*. 106, 103176 (2021).

186. Y. Wang, M. Zuo, H. Jin, M. Lai, J. Luo, Z. Cheng, Inhibition of ELF3 confers synthetic lethality of PARP inhibitor in non-small cell lung cancer. *J Recept Sig Transd*. 41, 304–311 (2021).

187. M. A. Boti, P. G. Adamopoulos, P. Tsiakanikas, A. Scorilas, Nanopore Sequencing Unveils Diverse Transcript Variants of the Epithelial Cell-Specific Transcription Factor Elf-3 in Human Malignancies. *Genes-basel*. 12, 839 (2021).

188. N. J. Goff, M. Brenière, C. J. Buehl, A. J. de Melo, H. Huskova, T. Ochi, T. L. Blundell, W. Mao, K. Yu, M. Modesti, K. Meek, Catalytically inactive DNA ligase IV promotes DNA repair in living cells. *Nucleic Acids Res*. 50, 11058–11071 (2022).

189. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, M. Imielinsk, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, A. P. C. G. Initiative, I. B. C. Consortium, I. M.-S. Consortium, I. PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, Signatures of mutational processes in human cancer. *Nature*. 500, 415–421 (2013).

190. H. E. Machado, E. Mitchell, N. F. Øbro, K. Kübler, M. Davies, D. Leongamornlert, A. Cull, F. Maura, M. A. Sanders, A. T. J. Cagan, C. McDonald, M. Belmonte, M. S. Shepherd, F. A. V. Braga, R. J. Osborne, K. Mahbubani, I. Martincorena, E. Laurenti, A. R. Green, G. Getz, P. Polak, K. Saeb-Parsy, D. J. Hodson, D. G. Kent, P. J. Campbell, Diverse mutational landscapes in human lymphocytes. *Nature*. 608, 724–732 (2022).

191. J. Ganz, L. J. Luquette, S. Bizzotto, C. L. Bohrson, H. Jin, M. B. Miller, Z. Zhou, A. Galor, P. J. Park, C. A. Walsh, Contrasting patterns of somatic mutations in neurons and glia reveal differential predisposition to disease in the aging human brain. *Biorxiv Prepr Serv Biology* (2023), doi:10.1101/2023.01.14.523958.

192. M. J. Rybin, M. Ramic, N. R. Ricciardi, P. Kapranov, C. Wahlestedt, Z. Zeier, Emerging Technologies for Genome-Wide Profiling of DNA Breakage. *Frontiers Genetics*. 11, 610386 (2021).

193. H. Li, J. Guo, G. Cheng, Y. Wei, S. Liu, Y. Qi, G. Wang, R. Xiao, W. Qi, W. Qiu, Identification and Validation of SNP-Containing Genes With Prognostic Value in Gastric Cancer via Integrated Bioinformatics Analysis. *Frontiers Oncol*. 11, 564296 (2021).

194. M. A. Slatter, A. R. Gennery, Update on DNA-Double Strand Break Repair Defects in Combined Primary Immunodeficiency. *Curr Allergy Asthm R*. 20, 57 (2020).

195. K. Felgentreff, S. N. Baxi, Y. N. Lee, K. Dobbs, L. A. Henderson, K. Csomos, E. N. Tsitsikov, M. Armanios, J. E. Walter, L. D. Notarangelo, Ligase-4 Deficiency Causes Distinctive Immune Abnormalities in Asymptomatic Individuals. *J Clin Immunol*. 36, 341–353 (2016).

196. P. Tuijnenburg, H. L. Allen, S. O. Burns, D. Greene, M. H. Jansen, E. Staples, J. Stephens, K. J. Carss, D. Biasci, H. Baxendale, M. Thomas, A. Chandra, S. Kiani-Alikhan, H. J. Longhurst, S. L. Seneviratne, E. Oksenhendler, I. Simeoni, G. J. de Bree, A. T. J. Tool, E. M. M. van Leeuwen, E. H. T. M. Ebberink, A. B. Meijer, S. Tuna, D. Whitehorn, M. Brown, E. Turro, A. J. Thrasher, K. G. C. Smith, J. E. Thaventhiran, T. W. Kuijpers, Z. Adhya, H. Alachkar, A. Anantharachagan, R. Antrobus, G. Arumugakani, C. Bacchelli, H. Baxendale, C. Bethune, S. Bibi, B. Boardman, C. Booth, M. Browning, M. Brownlie, S. Burns, A. Chandra, H. Clifford, N. Cooper, S. Davies, J. Dempster, L. Devlin, R. Doffinger, E. Drewe, D. Edgar, W. Egner, T. El-Shanawany, B. Gaspar, R. Ghurye, K. Gilmour, S. Goddard, P. Gordins, S. Grigoriadou, S. Hackett, R. Hague, L. Harper, G. Hayman, A. Herwadkar, S. Hughes, A. Huissoon, S. Jolles, J. Jones, P. Kelleher, N. Klein, T. Kuijpers, D. Kumararatne, J. Laffan, H. L. Allen, S. Lear, H. Longhurst, L. Lorenzo, J. Maimaris, A. Manson, E. McDermott, H. Millar, A. Mistry, V. Morrisson, S. Murng, I. Nasir, S. Nejentsev, S. Noorani, E. Oksenhendler, M. Ponsford, W. Qasim, E. Quinn, I. Quinti, A. Richter, C. Samarghitean, R. Sargur, S. Savic, S. Seneviratne, C. Sewall, F. Shackley, I. Simeoni, K. G. C. Smith, E. Staples, H. Stauss, C. Steele, J. Thaventhiran, M. Thomas, A. Thrasher, S. Welch, L. Willcocks, S. Workman, A. Worth, N. Yeatman, P. Yong, S. Ashford, J. Bradley, D. Fletcher, T. Hammerton, R. James, N. Kingston, W. Ouwehand, C. Penkett, F. L. Raymond, K. Stirrups, M. Veltman, T. Young, S. Ashford, M. Brown, N. Clements-Brod, J. Davis, E. Dewhurst, M. Erwood, A. Frary, R. Linger, J. Martin, S. Papadia, K. Rehnstrom, W. Astle, A. Attwood, M. Bleda, K. Carss, L. Daugherty, S. Deevi, S. Graf, D. Greene, C. Halmagyi, M. Haimel, F. Hu, R. James, H. L. Allen, V. Matser, S. Meacham, K. Megy, C. Penkett, O. Shamardina, K. Stirrups, C. Titterton, S. Tuna, E. Turro, P. Yu, J. von Ziegenweldt, A. Furnell, R. Mapeta, I. Simeoni, S. Staines, J. Stephens, K. Stirrups, D. Whitehorn, P. Rayner-Matthews, C. Watt, Loss-of-function nuclear factor κB subunit 1 (NFKB1) variants are the most common monogenic cause of common variable immunodeficiency in Europeans. *J Allergy Clin Immun*. 142, 1285–1296 (2018).

197. J. E. D. Thaventhiran, H. L. Allen, O. S. Burren, W. Rae, D. Greene, E. Staples, Z. Zhang, J. H. R. Farmery, I. Simeoni, E. Rivers, J. Maimaris, C. J. Penkett, J. Stephens, S. V. V. Deevi, A. Sanchis-Juan, N. S. Gleadall, M. J. Thomas, R. B. Sargur, P. Gordins, H. E. Baxendale, M. Brown, P. Tuijnenburg, A. Worth, S. Hanson, R. J. Linger, M. S. Buckland, P. J. Rayner-Matthews, K. C. Gilmour, C. Samarghitean, S. L. Seneviratne, D. M. Sansom, A. G. Lynch, K. Megy, E. Ellinghaus, D. Ellinghaus, S. F. Jorgensen, T. H. Karlsen, K. E. Stirrups, A. J. Cutler, D. S. Kumararatne, A. Chandra, J. D. M. Edgar, A. Herwadkar, N. Cooper, S. Grigoriadou, A. P. Huissoon, S. Goddard, S. Jolles, C. Schuetz, F. Boschann, S. Abbs, Z. Adhya, J. Adlard, M. Afzal, I. Ahmed, M. Ahmed, S. Ahmed, T. J. Aitman, H. Alachkar, J. Alamelu, R. Alikhan, C. E. Allen, L. Allen, D. J. Allsup, A. Alvi, G. Ambegaonkar, A. Anantharachagan, P. Ancliff, J. Anderson, R. Antrobus, R. Armstrong, G. Arno, G. Arumugakani, R. Arya, S. Ashford, W. J. Astle, A. Attwood, S. Austin, Y. Aydinok, W. Ayub, C. Babbs, C. Bacchelli, T.

Baglin, T. Bakchoul, T. K. Bariana, J. Barratt, J. Barwell, J. Baski, R. W. Bates, J. Batista, H. E. Baxendale, G. Baynam, D. L. Bennett, C. Bethune, N. Bhatnagar, S. Bibi, A. Bierzynska, T. Biss, M. A. K. Bitner-Glindzicz, M. Bleda, I. Blesneac, B. Boardman, P. Boddana, H. J. Bogaard, C. Booth, S. Boyce, J. R. Bradley, A. Brady, G. Breen, P. Brennan, C. Brewer, A. Briley, M. Brown, R. Brown, M. J. Browning, M. Brownlie, C. J. Bryson, R. J. Buchan, J. Buck, M. S. Buckland, T. Bueser, C. B. Diz, S. O. Burns, O. S. Burren, P. Calleja, J. Carmichael, G. Carr-White, K. J. Carss, R. Casey, E. Chalmers, J. Chambers, J. Chambers, M. M. Y. Chan, M. V. Chan, A. Chandra, F. Cheng, I. K. Chinn, P. F. Chinnery, M. Chitre, S. Chong, M. T. Christian, C. Church, E. M. Clement, N. C. Brod, H. Clifford, V. E. Clowes, G. Coghlan, E. Colby, T. R. P. Cole, J. H. Collins, P. W. Collins, R. Condliffe, H. Terence. Cook, S. Cook, V. Cookson, N. Cooper, P. A. Corris, A. Creaser-Myers, A. Crisp-Hihn, N. S. Curry, A. J. Cutler, R. D. Costa, C. Danesino, M. J. Daniels, D. Darby, L. C. Daugherty, E. G. Davies, S. Davies, J. Davis, G. J. de Bree, S. Deacock, P. B. Deegan, S. V. V. Deevi, J. Dempster, T. Dent, C. Deshpande, L. A. Devlin, E. F. Dewhurst, A. K. Dixit, P. H. Dixon, R. Doffinger, H. Dolling, N. Dormand, K. Downes, A. M. Drazyk, E. Drewe, D. Duarte, T. Dutt, J. D. M. Edgar, K. E. Edwards, W. Egner, M. N. Ekani, T. El-Shanawany, S. Elkhalifa, T. Elston, I. Emmerson, W. N. Erber, M. Erwood, M. C. Estiu, D. G. Evans, G. Evans, T. Everington, M. Eyries, J. H. R. Farmery, R. Favier, H. V. Firth, M. M. Fitzpatrick, D. Fletcher, F. A. Flinter, J. C. Fox, A. J. Frary, C. E. French, K. Freson, M. Frontini, B. Furie, D. P. Gale, H. J. Gall, A. Gardham, H. B. Gaspar, M. Gattens, N. Ghali, P. K. Ghataorhe, S. Ghio, H.-A. Ghofrani, R. Ghurye, J. S. R. Gibbs, R. D. Gilbert, K. C. Gilmour, B. Girerd, J. C. Girling, P. Gissen, N. S. Gleadall, S. Goddard, P. Gordins, K. M. Gorman, D. Gosal, S. Graf, L. Grassi, D. Greene, A. J. Greenhalgh, L. Greenhalgh, A. Greinacher, P. Gresele, P. G. Griffiths, S. Griffiths, S. Grigoriadou, D. Grozeva, S. J. Hackett, R. D. M. Hadden, C. Hadinnapola, R. Hague, W. M. Hague, M. Haimel, M. Hall, C. Halmagyi, T. Hammerton, H. L. Hanson, K. Harkness, A. R. Harper, L. Harper, C. Harris, C. Harrison, D. Hart, A. Hassan, G. Hayman, J. W. M. Heemskerk, S. Hegde, A. Henderson, R. H. Henderson, A. Hensiek, Y. M. C. Henskens, A. Herwadkar, J. Hodgson, J. Hoffman, S. Holden, M. Holder, R. Horvath, H. Houlden, A. C. Houweling, L. S. Howard, F. Hu, G. Hudson, S. Hughes, S. Hughes, A. E. H. in ‘t Veld, A. P. Huissoon, M. Humbert, M. E. Hurles, J. A. Hurst, V. Irvine, L. Izatt, R. James, P. Jeevaratnam, M. Johnson, S. A. Johnson, S. Jolles, J. D. Jolley, B. Jones, J. Jones, D. Josifova, N. Jurkute, Y. M. Karim, M. A. Karoshi, M. A. Kasanicki, H. Kazkaz, R. Kazmi, D. Keeling, P. Kelleher, A. M. Kelly, C. Kempster, F. Kennedy, S. Kiani, D. G. Kiely, N. Kingston, S. Kinsey, N. Klein, R. Klima, E. Knox, M. A. Kostadima, G. Kovacs, A. B. Koziell, R. Kreuzhuber, D. Krishnakumar, T. W. Kuijpers, A. Kumar, D. S. Kumararatne, M. A. Kurian, J. Laffan, M. A. Laffan, F. Lalloo, M. P. Lambert, H. L. Allen, S. H. A. Lawman, A. Lawrie, D. M. Layton, S. E. Lear, M. M. Lees, C. Lentaigne, A. P. Levine, A. J. P. Lewington, W. Li, R. Liesner, R. J. Linger, B. Liu, H. Longhurst, L. E. Lorenzo, E. Louka, S. L. Hadeler, P. A. Lyons, M. Macdougall, R. D. Machado, R. V. M. Ross, L. H. Mackillop, R. MacLaren, B. Madan, L. Magee, M. Mahdi-Rogers, E. R. Maher, J. Maimaris, M. Makris, S. Mangles, A. Manson, A. Manzur, R. Mapeta, K. J. Marchbank, P. B. Mark, S. Marks, H. S. Markus, H.-U. Marschall, A. Marshall, J. M. Martin, L. Masati, M. Mathias, V. Matser, E. L. Matthews, A. Maw, H. Maxwell, P. McAlinden, M. I. McCarthy, E. M. McDermott, S. J. McGowan, C. McJannet, H. McKinney, S. Meacham, A. J. Mead, I. M. Castello, S. Meehan, K. Megy, S. Mehta, C. L. Mercer, M. Michaelides, A. C.

Michell, D. Milford, C. M. Millar, H. Millar, A. Mistry, F. Moenen, S. Moledina, D. Montani, A. T. Moore, J. Moore, N. W. Morrell, V. Morrisson, M. Mozere, K. W. Muir, A. D. Mumford, S. H. K. Murng, I. Nasir, S. Nejentsev, M. Newnham, J. Ng, A. Ngoh, S. Noorani, M. Noori, P. Nurden, J. M. O'Sullivan, S. Obaji, S. Okoli, E. Oksenhendler, A. Olschewski, H. Olschewski, A. C. M. Ong, K. R. Ong, H. Oram, E. Ormondroyd, S. Othman, W. H. Ouwehand, A. Pantazis, S. Papadia, A. Papandreou, S.-M. Park, A. P. J. Parker, D. Parry, G. Parsons, K. John. Pasi, J. Paterson, J. H. Payne, A. J. Peacock, K. Peerlinck, C. J. Penkett, J. Pepke-Zaba, D. Perry, R. Petersen, B. Piechowski-Jozwiak, F. Pinto, G. J. Polwarth, M. J. Ponsford, S. Prasad, I. Prokopenko, B. Psaila, A. Pyle, W. Qasim, E. Quinn, I. Quinti, S. Raina, L. Ranganathan, J. Rankin, S. Rankin, A. Rao, F. L. Raymond, P. J. Rayner-Matthews, K. Rehnstrom, E. Reid, M. M. Reilly, T. Renton, S. Revel-Vilk, C. J. Rhodes, A. S. C. Rice, E. E. Richards, M. Richards, S. Richardson, A. Richter, L. Robert, I. Roberts, M. T. Rondina, E. Rosser, P. Rothwell, C. Roughley, N. B. Roy, K. Rue-Albrecht, O. Sadeghi-Alavijeh, M. A. Saleem, R. M. Salmon, N. J. Samani, C. Samarghitean, J. G. Sambrook, A. Sanchis-Juan, R. Sandford, S. Santra, R. B. Sargur, S. C. Satchell, S. Savic, L. Scelsi, G. Schotte, S. Schulman, H. Schulze, R. Scott, M. Scully, C. Searle, W. Seeger, S. L. Seneviratne, W. A. C. Sewell, D. Seyres, F. Shackley, O. Shamardina, S. E. Shapiro, P. Sharma, H. A. Shehata, D. Shipley, R. Shtoyerman, K. Sibson, L. Side, I. Simeoni, M. Simpson, M. C. Sims, M. D. Sinha, S. Sivapalaratnam, A.-B. Skytte, K. G. C. Smith, K. Snape, L. Sneddon, A. Sohal, F. Soubrier, L. Southgate, M. Southwood, M. Splitt, S. Staines, E. Staples, H. Stark, H. Stauss, C. L. Steele, D. Stein, P. E. Stein, J. Stephens, K. E. Stirrups, S. Stock, M. J. Stubbs, J. Suntharalingam, E. M. Swietlik, E. Symington, R. C. Tait, K. Talks, R. Y. Y. Tan, G. B. Taylor, J. Thachil, J. E. D. Thaventhiran, A. C. Themistocleous, D. C. Thomas, E. Thomas, M. J. Thomas, P. Thomas, D. A. Thompson, K. Thomson, A. J. Thrasher, C. Thys, T. Tilly, M. Tischkowitz, C. Titterton, J. A. Todd, C.-H. Toh, A. T. J. Tool, M. R. Toshner, M. Traylor, C. M. Treacy, P. Treadaway, R. C. Trembath, S. Trippier, S. Tuna, W. Turek, E. Turro, P. D. Upton, R. Urniaz, T. Vale, C. V. Geet, N. van Zuydam, A. M. Vandersteen, M. Vazquez-Lopez, M. W. M. Veltman, J. Vogt, J. von Ziegenweidt, A. V. Noordegraaf, A. Vora, M. J. A. Vries, E. L. Wakeling, N. Walker, S. M. Walker, R. Walsh, I. Wanjiku, J. S. Ware, T. Q. Warner, E. Wassmer, H. Watkins, H. G. Watson, C. Watt, D. Waugh, N. Webb, A. R. Webster, W. Wei, A. Welch, S. B. Welch, D. Werring, J. Wessels, S. K. Westbury, J.-P. W. Westwood, J. Wharton, D. Whitehorn, J. Whitworth, M. R. Wilkins, L. Willcocks, D. J. Williams, C. Williamson, E. K. S. Wong, N. Wood, Y. Wood, C. G. Woods, E. R. Woodward, S. Workman, S. J. Wort, A. Worth, K. Yates, N. Yeatman, P. F. K. Yong, T. Young, P. Yu, P. Yu-Wai-Man, E. Zlamalova, P. A. Lyons, M. E. Hurles, S. Savic, S. O. Burns, T. W. Kuijpers, E. Turro, W. H. Ouwehand, A. J. Thrasher, K. G. C. Smith, Whole-genome sequencing of a sporadic primary immunodeficiency cohort. *Nature*. 583, 90–95 (2020).

198. L. Hammarström, I. Vorechovsky, D. Webster, Selective IgA deficiency (SIgAD) and common variable immunodeficiency (CVID). *Clin Exp Immunol*. 120, 225–231 (2000).

199. L. D. Genova, S. Ceppi, M. Stefanelli, S. Esposito, IgA Deficiency and Nephrotic Syndrome in Children. *Int J Environ Res Pu*. 15, 1702 (2018).

200. C. Weemaes, I. Klasen, J. Göertz, M. Beldhuis-Valkis, O. Olafsson, A. Haraldsson, Development of Immunoglobulin A in Infancy and Childhood. *Scand J Immunol*. 58, 642–648 (2003).

201. J. Routes, M. Abinun, W. Al-Herz, J. Bustamante, A. Condino-Neto, M. T. D. L. Morena, A. Etzioni, E. Gambineri, E. Haddad, L. Kobrynski, F. L. Deist, S. Nonoyama, J. B. Oliveira, E. Perez, C. Picard, N. Rezaei, J. Sleasman, K. E. Sullivan, T. Torgerson, ICON: The Early Diagnosis of Congenital Immunodeficiencies. *J Clin Immunol*. 34, 398–424 (2014).

202. T. Al-Shaikhly, H. D. Ochs, Hyper IgE syndromes: clinical and molecular characteristics. *Immunol Cell Biol*. 97, 368–379 (2019).

203. A. Aghamohammadi, J. Mohammadi, N. Parvaneh, N. Rezaei, M. Moin, T. Espanol, L. Hammarstrom, Progression of Selective IgA Deficiency to Common Variable Immunodeficiency. *Int Arch Allergy Imm*. 147, 87–92 (2008).

204. S. K. Shapira, H. H. Jabara, C. P. Thienes, D. J. Ahern, D. Vercelli, H. J. Gould, R. S. Geha, Deletional switch recombination occurs in interleukin-4-induced isotype switching to IgE expression by human B cells. *Proc National Acad Sci*. 88, 7528–7532 (1991).

205. E. Vincendeau, W. Wei, X. Zhang, C. Planchais, W. Yu, H. Lenden-Hasse, T. Cokelaer, J. P. da Fonseca, H. Mouquet, D. J. Adams, F. W. Alt, S. P. Jackson, G. Balmus, C. Lescale, L. Deriano, SHLD1 is dispensable for 53BP1-dependent V(D)J recombination but critical for productive class switch recombination. *Nat Commun*. 13, 3707 (2022).

206. A. K. Ling, C. C. So, M. X. Le, A. Y. Chen, L. Hung, A. Martin, Double-stranded DNA break polarity skews repair pathway choice during intrachromosomal and interchromosomal recombination. *Proc National Acad Sci*. 115, 201720962 (2018).

207. H. IJspeert, P. A. van Schouwenburg, D. van Zessen, I. Pico-Knijnenburg, G. J. Driessen, A. P. Stubbs, M. van der Burg, Evaluation of the Antigen-Experienced B-Cell Receptor Repertoire in Healthy Children and Adults. *Front Immunol*. 7, 410 (2016).

208. N. Nouri, S. H. Kleinstein, Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data. *Plos Comput Biol*. 16, e1007977 (2020).

209. J. M. Lumsden, T. McCarty, L. K. Petiniot, R. Shen, C. Barlow, T. A. Wynn, H. C. Morse, P. J. Gearhart, A. Wynshaw-Boris, E. E. Max, R. J. Hodes, Immunoglobulin Class Switch Recombination Is Impaired in Atm-deficient Mice. *J Exp Medicine*. 200, 1111–1121 (2004).

210. S. Saito, R. Maeda, N. Adachi, Dual loss of human POLQ and LIG4 abolishes random integration. *Nat Commun*. 8, 16112 (2017).

211. S.-Y. Isobe, K. Nagao, N. Nozaki, H. Kimura, C. Obuse, Inhibition of RIF1 by SCAI Allows BRCA1-Mediated Repair. *Cell Reports*. 20, 297–307 (2017).

212. J. L. Crowe, X. S. Wang, Z. Shao, B. J. Lee, V. M. Estes, S. Zha, DNA-PKcs phosphorylation at the T2609 cluster alters the repair pathway choice during immunoglobulin class switch recombination. *Proc National Acad Sci*. 117, 22953–22961 (2020).

213. X. Xie, T. Gan, B. Rao, W. Zhang, R. A. Panchakshari, D. Yang, X. Ji, Y. Cao, F. W. Alt, F. Meng, J. Hu, C-terminal deletion-induced condensation sequesters AID from IgH targets in immunodeficiency. *Embo J*. 41, e109324 (2022).

214. X. Liu, T. Liu, Y. Shang, P. Dai, W. Zhang, B. J. Lee, M. Huang, D. Yang, Q. Wu, L. D. Liu, X. Zheng, B. O. Zhou, J. Dong, L.-S. Yeap, J. Hu, T. Xiao, S. Zha, R. Casellas, X. S. Liu, F.-L. Meng, ERCC6L2 promotes DNA orientation-specific recombination in mammalian cells. *Cell Res*. 30, 732–744 (2020).

215. Q. Pan-Hammarström, Y. Zhao, L. Hammarström, Class Switch Recombination: A Comparison Between Mouse and Human. *Adv Immunol*. 93, 1–61 (2007).

216. H. L. Borges, R. Linden, J. Y. Wang, DNA damage-induced cell death: lessons from the central nervous system. *Cell Res*. 18, 17–26 (2008).

217. N. P. Keegan, S. D. Wilton, S. Fletcher, Analysis of Pathogenic Pseudoexons Reveals Novel Mechanisms Driving Cryptic Splicing. *Frontiers Genetics*. 12, 806946 (2022).

218. S. Nagai, S. Sildever, N. Nishi, S. Tazawa, L. Basti, T. Kobayashi, Y. Ishino, Comparing PCR-generated artifacts of different polymerases for improved accuracy of DNA metabarcoding. *Metabarcoding Metagenomics*. 6, e77704 (2022).

219. J. Ouyang, Y. Zhong, Y. Zhang, L. Yang, P. Wu, X. Hou, F. Xiong, X. Li, S. Zhang, Z. Gong, Y. He, Y. Tang, W. Zhang, B. Xiang, M. Zhou, J. Ma, Y. Li, G. Li, Z. Zeng, C. Guo, W. Xiong, Long non-coding RNAs are involved in alternative splicing and promote cancer progression. *Brit J Cancer*. 126, 1113–1124 (2022).

220. J. O'Brien, H. Hayder, Y. Zayed, C. Peng, Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front Endocrinol*. 9, 402 (2018).

221. A. Kozomara, M. Birgaoanu, S. Griffiths-Jones, miRBase: from microRNA sequences to function. *Nucleic Acids Res*. 47, gky1141- (2018).

222. B. H. R. D. Fonseca, D. S. Domingues, A. R. Paschoal, mirtronDB: a mirtron knowledge base. *Bioinformatics*. 35, btz153 (2019).

223. P. P. Amaral, M. B. Clark, D. K. Gascoigne, M. E. Dinger, J. S. Mattick, lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res*. 39, D146–D151 (2011).

224. P. J. Hilleren, R. Parker, Cytoplasmic Degradation of Splice-Defective Pre-mRNAs and Intermediates. *Mol Cell*. 12, 1453–1465 (2003).

225. L. K. Lerner, D. Bonte, M. L. Guillou, M. M. Mohammad, Z. Kasraian, A. Sarasin, E. Despras, S. Aoufouchi, Expression of Constitutive Fusion of Ubiquitin to PCNA Restores the Level of Immunoglobulin A/T Mutations During Somatic Hypermutation in the Ramos Cell Line. *Front Immunol*. 13, 871766 (2022).

226. A. J. Matthews, S. Zheng, L. J. DiMenna, J. Chaudhuri, Chapter One Regulation of Immunoglobulin Class-Switch Recombination Choreography of Noncoding Transcription, Targeted DNA Deamination, and Long-Range DNA Repair. *Adv Immunol*. 122, 1–57 (2014).

227. A. Kakarougkas, A. Ismail, K. Klement, A. A. Goodarzi, S. Conrad, R. Freire, A. Shibata, M. Lobrich, P. A. Jeggo, Opposing roles for 53BP1 during homologous recombination. *Nucleic Acids Res*. 41, 9719–9731 (2013).

228. H. Dev, T.-W. W. Chiang, C. Lescale, I. de Krijger, A. G. Martin, D. Pilger, J. Coates, M. Sczaniecka-Clift, W. Wei, M. Ostermaier, M. Herzog, J. Lam, A. Shea, M. Demir, Q. Wu, F. Yang, B. Fu, Z. Lai, G. Balmus, R. Belotserkovskaya, V. Serra, M. J. O'Connor, A. Bruna, P. Beli, L. Pellegrini, C. Caldas, L. Deriano, J. J. L. Jacobs, Y. Galanty, S. P. Jackson, Shieldin complex promotes DNA end-joining and counters homologous recombination in BRCA1-null cells. *Nat Cell Biol*. 20, 954–965 (2018).

229. A. Yamane, D. F. Robbiani, W. Resch, A. Bothmer, H. Nakahashi, T. Oliveira, P. C. Rommel, E. J. Brown, A. Nussenzweig, M. C. Nussenzweig, R. Casellas, RPA Accumulation during Class Switch Recombination Represents 5′–3′ DNA-End Resection during the S–G2/M Phase of the Cell Cycle. *Cell Reports*. 3, 138–147 (2013).

230. A. Björkman, P. Qvist, L. Du, M. Bartish, A. Zaravinos, K. Georgiou, A. D. Børglum, R. A. Gatti, T. Törngren, Q. Pan-Hammarström, Aberrant recombination and repair during immunoglobulin class switching in BRCA1-deficient human B cells. *Proc National Acad Sci*. 112, 2157–2162 (2015).

231. A. R. Chaudhuri, E. Callen, X. Ding, E. Gogola, A. A. Duarte, J.-E. Lee, N. Wong, V. Lafarga, J. A. Calvo, N. J. Panzarino, S. John, A. Day, A. V. Crespo, B. Shen, L. M. Starnes, J. R. de Ruiter, J. A. Daniel, P. A. Konstantinopoulos, D. Cortez, S. B. Cantor, O. Fernandez-Capetillo, K. Ge, J. Jonkers, S. Rottenberg, S. K. Sharan, A. Nussenzweig, Replication fork stability confers chemoresistance in BRCA-deficient cells. *Nature*. 535, 382–387 (2016).

232. J. R. Chapman, P. Barral, J.-B. Vannier, V. Borel, M. Steger, A. Tomas-Loba, A. A. Sartori, I. R. Adams, F. D. Batista, S. J. Boulton, RIF1 Is Essential for 53BP1-Dependent Nonhomologous End Joining and Suppression of DNA Double-Strand Break Resection. *Mol Cell*. 49, 858–871 (2013).

233. A. Bothmer, D. F. Robbiani, N. Feldhahn, A. Gazumyan, A. Nussenzweig, M. C. Nussenzweig, 53BP1 regulates DNA resection and the choice between classical and alternative end joining during class switch recombination. *J Exp Med*. 207, 855–865 (2010).

234. A. Cruz-García, A. López-Saavedra, P. Huertas, BRCA1 Accelerates CtIP-Mediated DNA-End Resection. *Cell Reports*. 9, 451–459 (2014).

235. D. Menolfi, W. Jiang, B. J. Lee, T. Moiseeva, Z. Shao, V. Estes, M. G. Frattini, C. J. Bakkenist, S. Zha, Kinase-dead ATR differs from ATR loss by limiting the dynamic exchange of ATR and RPA. *Nat Commun*. 9, 5351 (2018).

236. T. C. Nepomuceno, G. D. Gregoriis, F. M. B. de Oliveira, G. Suarez-Kurtz, A. N. Monteiro, M. A. Carvalho, The Role of PALB2 in the DNA Damage Response and Cancer Predisposition. *Int J Mol Sci*. 18, 1886 (2017).

237. A. Piazza, H. Bordelet, A. Dumont, A. Thierry, J. Savocco, F. Girard, R. Koszul, Cohesin regulates homology search during recombinational DNA repair. *Nat Cell Biol*. 23, 1176–1186 (2021).

238. J. M. Daley, P. Sung, 53BP1, BRCA1, and the Choice between Recombination and End Joining at DNA Double-Strand Breaks. *Mol Cell Biol*. 34, 1380–1388 (2014).

239. K. Kumazaki, B. Tirosh, R. Maehr, M. Boes, T. Honjo, H. L. Ploegh, AID−/−µs−/− Mice Are Agammaglobulinemic and Fail to Maintain B220−CD138+ Plasma Cells. *J Immunol*. 178, 2192–2203 (2007).

240. B. Schwer, P.-C. Wei, A. N. Chang, J. Kao, Z. Du, R. M. Meyers, F. W. Alt, Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. *Proc National Acad Sci*. 113, 2258–2263 (2016).

241. N. Michel, H. M. R. Young, N. D. Atkin, U. Arshad, R. Al-Humadi, S. Singh, A. Manukyan, L. Gore, I. E. Burbulis, Y.-H. Wang, M. J. McConnell, Transcription-associated DNA DSBs activate p53 during hiPSC-based neurogenesis. *Sci Rep-uk*. 12, 12156 (2022).

242. I. Rybanska-Spaeder, T. L. Reynolds, J. Chou, M. Prakash, T. Jefferson, D. L. Huso, S. Desiderio, S. Franco, 53BP1 Is Limiting for NHEJ Repair in ATM-deficient Model Systems That Are Subjected to Oncogenic Stress or Radiation. *Mol Cancer Res*. 11, 1223–1234 (2013).

243. G. Balmus, D. Pilger, J. Coates, M. Demir, M. Sczaniecka-Clift, A. C. Barros, M. Woods, B. Fu, F. Yang, E. Chen, M. Ostermaier, T. Stankovic, H. Ponstingl, M. Herzog, K. Yusa, F. M. Martinez, S. T. Durant, Y. Galanty, P. Beli, D. J. Adams, A. Bradley, E. Metzakopian, J. V. Forment, S. P. Jackson, ATM orchestrates the DNA-damage response to counter toxic non-homologous end-joining at broken replication forks. *Nat Commun*. 10, 87 (2019).

244. A. T. S. Boone, I. K. Chinn, C. Alaez-Versón, M. A. Yamazaki-Nakashimada, K. Carrillo-Sánchez, M. de la L. H. García-Cruz, M. C. Poli, M. E. G. Serrano, E. A. M. Torres, D. M. Zermeño, L. R. Forbes, F. J. Espinosa-Rosales, S. E. Espinosa-Padilla, J. S. Orange, S. O. L. Reyes, Failing to Make Ends Meet: The Broad Clinical Spectrum of DNA Ligase IV Deficiency. Case Series and Review of the Literature. *Frontiers Pediatrics*. 6, 426 (2019).

245. C. Scheepers, V. Bekker, C. Anthony, S. I. Richardson, B. Oosthuysen, T. Moyo, P. Kgagudi, D. Kitchin, M. Nonyane, T. York, D. Mielke, B. M. Mabvakure, Z. Sheng, B. E. Lambson, A. Ismail, N. J. Garrett, S. S. A. Karim, L. Shapiro, C. Williamson, L. Morris, P. L. Moore, Antibody Isotype Switching as a Mechanism to Counter HIV Neutralization Escape. *Cell Reports*. 33, 108430 (2020).

246. K. Kitaura, H. Yamashita, H. Ayabe, T. Shini, T. Matsutani, R. Suzuki, Different Somatic Hypermutation Levels among Antibody Subclasses Disclosed by a New Next-Generation Sequencing-Based Antibody Repertoire Analysis. *Front Immunol*. 8, 389 (2017).

247. A. Tinguely, G. Chemin, S. Péron, C. Sirac, S. Reynaud, M. Cogné, L. Delpy, Cross Talk between Immunoglobulin Heavy-Chain Transcription and RNA Surveillance during B Cell Development. *Mol Cell Biol*. 32, 107–117 (2012).

248. D. Lindholm, L. Korhonen, O. Eriksson, S. Kõks, Recent Insights into the Role of Unfolded Protein Response in ER Stress in Health and Disease. *Frontiers Cell Dev Biology*. 5, 48 (2017).

249. S. Camero, S. Ceccarelli, F. D. Felice, F. Marampon, O. Mannarino, L. Camicia, E. Vescarelli, P. Pontecorvi, B. Pizer, R. Shukla, A. Schiavetti, M. G. Mollace, A. Pizzuti, V. Tombolini, C. Marchese, F. Megiorni, C. Dominici, PARP inhibitors affect growth, survival and radiation susceptibility of human alveolar and embryonal rhabdomyosarcoma cell lines. *J Cancer Res Clin*. 145, 137–152 (2019).

250. P. Gralewska, A. Gajek, A. Marczak, A. Rogalska, Metformin Affects Olaparib Sensitivity through Induction of Apoptosis in Epithelial Ovarian Cancer Cell Lines. *Int J Mol Sci*. 22, 10557 (2021).

251. E. K. Lee, U. A. Matulonis, PARP Inhibitor Resistance Mechanisms and Implications for Post-Progression Combination Therapies. *Cancers*. 12, 2054 (2020).

252. E. M. Cortizas, A. Zahn, S. Safavi, J. A. Reed, F. Vega, J. M. D. Noia, R. E. Verdun, UNG protects B cells from AID-induced telomere lossTelomere protection in B cells. *J Exp Medicine*. 213, 2459–2472 (2016).

253. O. Dézé, B. Laffleur, M. Cogné, Roles of G4-DNA and G4-RNA in Class Switch Recombination and Additional Regulations in B-Lymphocytes. *Molecules*. 28, 1159 (2023).

254. Q. Qiao, L. Wang, F.-L. Meng, J. K. Hwang, F. W. Alt, H. Wu, AID Recognizes Structured DNA for Class Switch Recombination. *Mol Cell*. 67, 361-373.e4 (2017).

255. J. Hu, R. M. Meyers, J. Dong, R. A. Panchakshari, F. W. Alt, R. L. Frock, Detecting DNA double-stranded breaks in mammalian genomes by linear amplification–mediated high-throughput genome-wide translocation sequencing. *Nat Protoc*. 11, 853–871 (2016).

256. S. G. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj, M. F. Polz, PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Appl Environ Microb*. 71, 8966–8969 (2005).

257. J. Bruijnesteijn, M. van der Wiel, N. G. de Groot, R. E. Bontrop, Rapid Characterization of Complex Killer Cell Immunoglobulin-Like Receptor (KIR) Regions Using Cas9 Enrichment and Nanopore Sequencing. *Front Immunol*. 12, 722181 (2021).

258. L. E. Brace, S. C. Vose, D. F. Vargas, S. Zhao, X. Wang, J. R. Mitchell, Lifespan extension by dietary intervention in a mouse model of Cockayne Syndrome uncouples early postnatal development from segmental progeria. *Aging Cell*. 12, 1144–1147 (2013).

259. T. Shih, S. De, B. J. Barnes, RNAi Transfection Optimized in Primary Naïve B Cells for the Targeted Analysis of Human Plasma Cell Differentiation. *Front Immunol*. 10, 1652 (2019).

260. T. Nojima, K. Haniuda, T. Moutai, M. Matsudaira, S. Mizokawa, I. Shiratori, T. Azuma, D. Kitamura, In-vitro derived germinal centre B cells differentially generate memory B or plasma cells in vivo. *Nat Commun*. 2, 465 (2011).

10. AFFIDAVIT

Statutory Declaration

“I, Clara Vázquez García, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic “A novel methodology to identify DNA repair deficiencies by means of genomic scars generated in the course of antibody diversification” or “Eine neuartige Methode zur Identifizierung von DNA-Reparaturdefiziten anhand von genomischen Narben, die im Verlauf der Antikörperdiversifizierung auftreten”, independently and without the support of third parties, and that I used no other sources and aids than those stated.

All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding practical work, laboratory regulations, statistical processing) and results (in particular regarding figures, charts and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me.”

30.06.2024 _____

Date Signature

11. CURRICULUM VITAE

My curriculum vitae does not appear in the electronic version of my paper for reasons of data protection.

12. PUBLICATION LIST

M. Lebedin, **C. V. García**, L. Spatt, C. Ratswohl, C. Thibeault, L. Ostendorf, T. Alexander, F. Paul, L. E. Sander, F. Kurth, K. Rosa, Discriminating promiscuous from target-specific autoantibodies in COVID-19. Eur. J. Immunol., 2250210 (2023).

M. Lebedin, M. Foglierini, S. Khorkova, **C. V. García**, C. Ratswohl, A. N. Davydov, M. A. Turchaninova, C. Daubenberger, D. M. Chudakov, A. Lanzavecchia, K. de la Rosa, Different classes of genomic inserts contribute to human antibody diversity. Proc National Acad Sci. 119 (2022).

13. ACKNOWLEDGMENTS

I would like to thank my mentor and supervisor, Kathrin de la Rosa. She guided me all the way through my thesis and was a crucial turning point in the decision-making of my project, especially in developing the tool to identify DNA repair malfunction. She is the mentor everyone would like to have, who has time for you and does her best to make you succeed. I would also like to thank Benedikt Obermayer, who implemented the bioinformatical expertise to develop SWIBRID, met with me weekly to refine it and still works with us to develop the ultimate tool to diagnose DNA repair deficiencies. Thank you Kathrin and Benedikt.

I would like to thank my supervisors, Michela Di Virgilio and Uta Höpken, with whom I had great discussions in the TAC meetings and who helped me by offering the CH12 cell lines and the usage of the Bioanalyzer.

I very much appreciate the kindness of Qiang Pan-Hammerström, who contributed to developing my thesis and identifying DNA repair malfunction by supplying human DNA repair deficient samples.

The AG de la Rosa team was a great support, especially Mikhail Lebedin, with whom I had great scientific discussions and who helped me generate the VDJ repertoire. And also to Cathrin and Lisa, who helped push the experiments to be done on time. To the team in general for their input in the lab meetings and the great time together.

Last, to my family and friends. Hannes, who helped me shape the Zusammenfassung, gave me feedback on the thesis and dealt with me at home during this journey's difficult moments. Mamá and Papá, who were always willing to help me from the other line of the phone. My sisters Irene and María, who were always trying to understand what I was doing. And my friends from Berlin, especially Cristian and Marta, with whom I could have a good time not thinking about the thesis, which was equally important to make this journey happen.

Thank you all.

14. CERTIFICATE OF STATISTICIAN

CHARITÉ

CharitéCentrum für Human- und Gesundheitswissenschaften

Charité | Campus Charité Mitte | 10117 Berlin

Institut für Biometrie und klinische Epidemiologie (iBikE)
Direktor: Prof. Dr. Frank Konietschke
Postanschrift:
Charitéplatz 1 | 10117 Berlin
Besucheranschrift:
Reinhardtstr. 58 | 10117 Berlin
Tel. +49 (0)30 450 562171
frank.konietschke@charite.de
https://biometrie.charite.de/

Name, Vorname: Vazquez Garcia, Clara
Emailadresse: clara.vazquez-garcia@charite.de
Matrikelnummer: 226493
PromotionsbetreuerIn: Prof. Kathrin de la Rosa
Promotionsinstitution / Klinik: Max Delbrück Center for Molecular Medicine

Bescheinigung

Hiermit bescheinige ich, dass Frau *Clara Vazquez Garcia* innerhalb der Service Unit Biometrie des Instituts für Biometrie und klinische Epidemiologie (iBikE) bei mir eine statistische Beratung zu einem Promotionsvorhaben wahrgenommen hat. Folgende Beratungstermine wurden wahrgenommen:

- Termin 1: 09.03.2023

Folgende wesentliche Ratschläge hinsichtlich einer sinnvollen Auswertung und Interpretation der Daten wurden während der Beratung erteilt:

- Prüfung der Normalverteilung
- Die lineare Regression
- Einfaktorielle Varianzanalyse (ANOVA) und Posthoc-Analyse
- Der t-Test für unabhängige Stichproben
- Nonparametrische Verfahren: Wilcoxon signed-rank Test und Kruskal-Wallis Test
- Principal component analysis (PCA)

Diese Bescheinigung garantiert nicht die richtige Umsetzung der in der Beratung gemachten Vorschläge, die korrekte Durchführung der empfohlenen statistischen Verfahren und die richtige Darstellung und Interpretation der Ergebnisse. Die Verantwortung hierfür obliegt allein dem Promovierenden. Das Institut für Biometrie und klinische Epidemiologie übernimmt hierfür keine Haftung.

Datum: 09.03.2023

Name der Beraterin: Pimrapat Gebert

Unterschrift Beraterin, Institutsstempel

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN
Institut für Biometrie und
Klinische Epidemiologie
Campus Charité Mitte
Charitéplatz 1 | D-10117 Berlin
Sitz: Reinhardtstr. 58