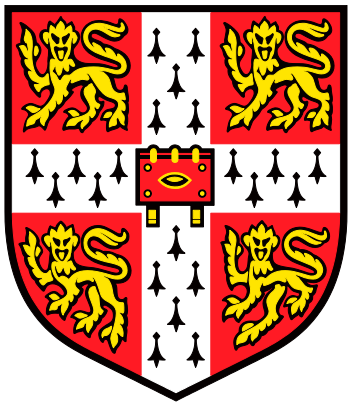
STUDIES OF B CELL DEVELOPMENT AND V(D)J RECOMBINATION

Peter Chovanec

Sidney Sussex College



University of Cambridge
Babraham Institute

This dissertation is submitted for the degree of doctor of philosophy

September 2018

Declaration

I hereby declare that my thesis entitled 'Studies Of B Cell Development And V(D)J Recombination' is the result of my own work except that which has been done in collaboration, as stated in the table of acknowledgements and in the text. This work has not been previously submitted for a degree, diploma or other qualification.

This thesis does not exceed the word limit of 60 000 words.

STUDIES OF B CELL DEVELOPMENT AND V(D)J RECOMBINATION

Peter Chovanec

Abstract

The process of generating the vast diversity of immunoglobulin receptors and secreted antibodies begins with the recombination of the joining (J_H), diversity (D_H) and variable (V_H) genes in the immunoglobulin heavy chain locus. The ability to produce antibodies is restricted to the B cell lineage and is tightly regulated, starting with the temporal separation of the recombination process, in which D_H-J_H precedes V_H-D_HJ_H recombination. Successful recombination of both heavy and light chain loci results in the expression of an antigen receptor on the cell surface. Subsequent selection stages remove non-functional and autoreactivity receptors from the final pool of antigen responding B cells that ultimately give rise to antibody secreting plasma cells. Understanding the complexity of the recombination processes and the diversity of the resulting antibody repertoire has been a major focus of academic and industrial research alike.

Therapeutic monoclonal antibodies have seen many successful applications within the clinic and they constitute a billion-dollar industry. However, limitations therein have resulted in the emergence of antibody engineering approaches and the use of natural sources of alternative heavy chain only antibodies (HCAbs/nanobodies). The biotechnology company Crescendo Biologics has taken the highly desired characteristics of HCAbs a step further with the creation of a mouse platform capable of producing fully humanized HCAbs. The Crescendo platform presents a unique opportunity to expand our understanding of how mouse B cell development functions by exploiting the features of heavy chain only antibody production. Furthermore, the platform enables the expansion of our limited knowledge of the epigenetic mechanisms involved in the recombination of the human immunoglobulin heavy chain locus.

Using flow cytometry, with dimensionality reduction analysis approaches, I investigated B cell development in the context of HCAbs. These studies revealed a previously uncharacterised developmentally intermediate B cell population. Due to ethical and availability limitations to studies of human bone marrow, the primary pre-selection human B cell repertoire has not been studied in detail. The isolation of several B cell developmental stages and the use of our novel DNA-based high-throughput unbiased repertoire quantification technique, VDJ-seq, allowed me to study recombination of the human IGH locus sequence and observe HCAb repertoire selection within the mouse environment.

The adaptation of next generation sequencing techniques to antigen receptor repertoire quantification has provided an unprecedented insight into repertoire diversity and the alterations it undergoes during infection or ageing. Our VDJ-seq assay is unique in its ability to interrogate DNA recombinants. To expand its capabilities, I investigated several limitations of the technique, including mispriming and PCR/sequencing errors, and implemented experimental and bioinformatics solutions to overcome them, which included the creation of a comprehensive analysis workflow.

Finally, I have developed and applied a novel network visualisation method for genome-wide promoter interaction data generated by promoter capture Hi-C. The availability of high quality human pluripotent stem cell datasets allowed me to utilise the new techniques to further our understanding of the dynamics of genome organisation during early human embryonic development. This visualisation approach will be directly applicable to understanding B cell development.

Acknowledgments

There are many people I would like to thank for their support during my PhD. Firstly, my supervisors Anne Corcoran and Colette Johnston for their incredible support, guidance and helpful advice.

I would also like to thank many members of the Babraham Institute, especially those within the Nuclear Dynamics and Lymphocyte Signalling and Development ISPs and the core facilities.

I would also like to give thanks to all past and present members of the Corcoran lab for making it an enjoyable experience, especially Ola for all her wonderful baking. Likewise, I would also like to thank Peter Rugg-Gunn and Stefan Schoenfelder for their insightful discussions and valuable mentorship.

Further thanks go to all of my fantastic collaborators who have helped to make this work possible.

Hlavne by som chcel poďakovať mamine a tatinovi. Bez ich nekonečnej podpory by som nikdy neuspel.

Finally, I am eternally grateful to all of my family and my wonderful partner for their endless love and support and for sticking by me through thick and thin. Without them, I would never have accomplished as much as I have. I love you all.

Table of Acknowledgements

Initial training in techniques and laboratory practice and subsequent mentoring:

Dr Anne Corcoran
Dr Colette Johnston
Miss Amanda Jayne Collier
Dr Daniel Bolland
Dr Jannek Hauser
Dr Amanda Baizan-Edge
Dr Bryony Stubbs
Dr Olga Mielczarek
Dr. Peter Rugg-Gunn
Dr. Stefan Schoenfelder
Dr Louise Matheson
Dr Jon Houseley
Dr Ryan Hull
Dr Klaus Okkenhaug
Dr Suzanne Turner
Dr Simon Andrews

Data obtained from a technical service provider:

Dr Simon Andrews - Bioinformatics training
Dr Felix Krueger - Bioinformatics training, original VDJ-seq pipeline developer
Dr Rachael Walker - Flow Cytometry Core training & FACS
Arthur Davis - FACS
Dr Simon Walker - Imaging facility training
Miss Kristina Tabbada – Sequencing facility
Sanger sequencing - Beckman Coulter
Oligonucleotide synthesis - Sigma-Aldrich and IDT
Dr. Steven Wingett – Processing data through HiCUP and CHiCAGO

Data produced jointly:

Bone marrow extraction from femurs and tibias:
Dr Daniel Bolland
Dr Amanda Baizan-Edge
Dr Bryony Stubbs
Dr Olga Mielczarek

Scientific and conceptual insight into PCHiC analysis of naïve and primed hPSCs:
Dr. Peter Rugg-Gunn
Dr. Amanda Jayne Collier
Dr. Stefan Schoenfelder
Dr. Anne Corcoan

Naïve and primed hPSC data generation:
Dr. Amanda Jayne Collier
Dr. Stefan Schoenfelder

Data/materials provided by someone else:

Dr Louise Matheson – Human PBMC VDJ-seq data
Dr Daniel Bolland – Wildtype VDJ-seq data
Dr Yumin Teng – Transgene information and YAC construct
Mouse husbandry - Babraham Biological Services Unit staff and Charles River
Dr Felix Krueger – Quality trimming of fastq reads
Mr Sam Rees – Wildtype VDJ-seq data

Table of contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

CTG - Crescendo transgenic mouse
ELP – early lymphoid progenitors
CLP – common lymphoid progenitors
RAG – recombination-activating gene
RSS – Recombination signal sequence
SJ - signal joint
CJ – coding joint
V – variable
D – diversity
J - joining
RIC – recombination information content
SLC – surrogate light chain
NHEJ – non-homologous end joining
LC – light chain (protein)
HC – heavy chain (protein)
FOXO – forkhead box O
CDKs - cyclin-dependent kinases
HCD – heavy chain disease
ATM - ataxia-telangiectasia mutated
PCH - pericentromeric heterochromatin
IL-7R - interleukin-7 receptor
Igl – immunoglobulin light chain loci
Igh – immunoglobulin heavy chain locus
iEκ - Igκ intronic enhancer
ncRNA - non-coding RNA
lncRNA – long non-coding RNA
FO – follicular
MZ – marginal zone
SHM – somatic hypermutation
AID - activation-induced cytidine deaminase
CSR - class switch recombination
mAbs - monoclonal antibodies
HAMA - human anti-mouse antibodies
dAbs - single domain antibodies
HCAbs - heavy chain only antibodies
VHH - camelid variable domains
VH - human variable domain
WT- wildtype C57Bl/6JBabr mice
TGN - *trans*-Golgi network
PtC - phosphatidylcholine
ANA - anti-nuclear antibodies
sIgM – secreted IgM
intEF - intermediate EF
SOM - self-organising map
MST - minimal spanning tree
TKO – triple knockout
RBs - Russell bodies
SSC-A - side scatter area
YAC - yeast artificial chromosome
TLA - targeted locus amplification
3C - chromosome conformation capture
4C – circularised chromosome conformation capture
ORF - open reading frame
Del1-4 – deletion 1-4
aa – amino acid
UMI – unique molecular identifier
CV – coefficient of variation
MSA – multiple sequence alignment
CDR – complementarity-determining regions
HMM - hidden Markov model

LIST OF PUBLICATIONS

The following publications are the result of work performed during my PhD:

Koohy, H.*, Bolland, D.J.*, Matheson, L.S.*, Schoenfelder, S., Stellato, C., Dimond, A., Várnai, C., **Chovanec, P.**, Chessa, T., Denizot, J., Garcia, R.M., Wingett, S.W., Freire-Pritchett, P., Nagano, T., Mielczarek, O., Baizan-Edge, A., Stubbs, B.A., Hawkins, P., Stephens, L., Elderkin, S., Spivakov, M., Fraser, P., Corcoran, A.E., and Varga-Weisz, P.D. (2018) Genome organisation and chromatin analysis identifies transcriptional downregulation of insulin-like growth factor signalling as a hallmark of ageing in developing B cells. Genome Biol. *19*, 126.

Chovanec, P., Bolland, D.J., Matheson, L.S., Wood, A.L., Corcoran, A.E. (2018). Unbiased quantification of immunoglobulin diversity at the DNA level with VDJ-seq. Nat. Protoc. *13*, 1232–1252.

Collier, A.J.†, Panula, S.P.†, **Chovanec, P.***, Schell, J.P.*, Reyes, A.P., Petropoulos, S., Corcoran, A.E., Walker, R., Douagi, I., Lanner, F., Rugg-Gunn, P.J. (2017). Comprehensive Cell Surface Protein Profiling Identifies Specific Markers of Human Naive and Primed Pluripotent States. Cell Stem Cell *20*, 874–890.e7

Matheson, L.S., Bolland, D.J., **Chovanec, P.**, Krueger, F., Andrews, S., Koohy, H., and Corcoran, A. (2017). Local chromatin features including PU.1 and IKAROS binding and H3K4 methylation shape the repertoire of immunoglobulin kappa genes chosen for V(D)J recombination. Front. Immunol. *8*, 1550.

Bolland, D.J., Koohy, H., Wood, A.L., Matheson, L.S., Krueger, F., Stubbington, M.J.T., Baizan-Edge, A., **Chovanec, P.**, Stubbs, B.A., Tabbada, K., Andrews, S.R., Spivakov, M., Corcoran, A.E. (2016). Two Mutually Exclusive Local Chromatin States Drive Efficient V(D)J Recombination. Cell Rep. *15*, 2475–2487.

Equal contributions: *; Co-first author: †

1 Introduction

During a vertebrate's life, an enormous variety of pathogens are encountered on a daily basis ranging from bacteria to viruses and other parasitic organisms. The adaptive branch of the immune system represents one of the main defences against these invaders. It is uniquely capable of tailoring the immune response to the encountered pathogens and forming a memory of past encounters for a swift future defence. This unique ability is what allows vaccines to train our immune system to protect us from otherwise deadly pathogens. The adaptive ability stems from the vast diversity of immunoglobulins (antibodies when secreted) produced by the B cell and T cell lymphocyte lineages that are capable of binding to pathogen epitopes with high affinity and alerting the immune system of the invader's presence. The diversity of immunoglobulins stems from the combinatorial recombination of three gene segments, which together create the variable exon that forms the epitope recognition site of immunoglobulins. Further diversification mechanisms in B cells select and hone low-affinity immunoglobulins into high-affinity pathogen specific ones, which are eventually secreted in the form of antibodies.

The vast majority of our knowledge about the immunoglobulin recombination process comes from studies in mice. During my PhD, I focused on developing methods for the quantification of immunoglobulin repertoires and examining the recombination of a human immunoglobulin transgene model, with the goal of further understanding the processes that govern recombination in humans. The human transgene model is a unique mouse therapeutic platform capable of producing high-affinity heavy chain only antibodies. I begin by introducing B cell development within a mouse and examining the different selection checkpoints each B cell must pass through. For a mouse to produce heavy chain only antibodies several engineering feats had to be accomplished, the consequence of which I examine in terms of B cell development. I later examine the transgene and attempt to correlate the proximity of epigenetic factors with recombination frequency. I next describe the optimisation of an unbiased repertoire quantification assay developed in the lab and the development of a comprehensive analysis pipeline. Finally, I will delve into a side-project involving the development of a new visualisation method for promoter capture Hi-C and its use in examining changes in a model of early human embryonic development.

1.1 Initiation of early B cell development within the bone marrow

All blood-cell lineages originate from self-renewing hematopoietic stems cells (HSCs) that reside within the bone marrow. As they progress down different developmental lineages, they acquire lineage specific characteristics. The earliest indication of differentiation down the B cell lineage are presented by the transcription of terminal deoxynucleotidyl transferase (TdT), which is responsible for non-template (N) nucleotide addition during antigen receptor recombination. The expression of TdT is

thought to be restricted to the B cell lineage. Collectively, the HSCs that are biased towards the lymphoid lineage are known as early lymphoid progenitors (ELP) (Figure 1-1) (Hardy et al., 2007). In a subset of ELPs, further progression down the B cell lineage into the common lymphoid progenitors (CLP) is accompanied by the expression and translation of the recombination-activating genes 1 and 2 (RAG1/2). The RAG1/2 recombinases initiate D_H-J_H recombination of the immunoglobulin heavy chain (Igh), which leads to the loss of differentiation potential into non-lymphoid lineages (Igarashi et al., 2002). The B cell lineage specification is accompanied by the expression and repression of multiple transcription factors (TFs) that regulate the developmental progression. Both ELPs and CLPs still retain multi-lineage potential (Busslinger, 2004), and plasticity of the B cell lineage is not lost until the expression of the B220 and CD19 cell surface proteins in progenitor B cells (pro-B; fraction B; Figure 1-1) (Rumfelt et al., 2006).

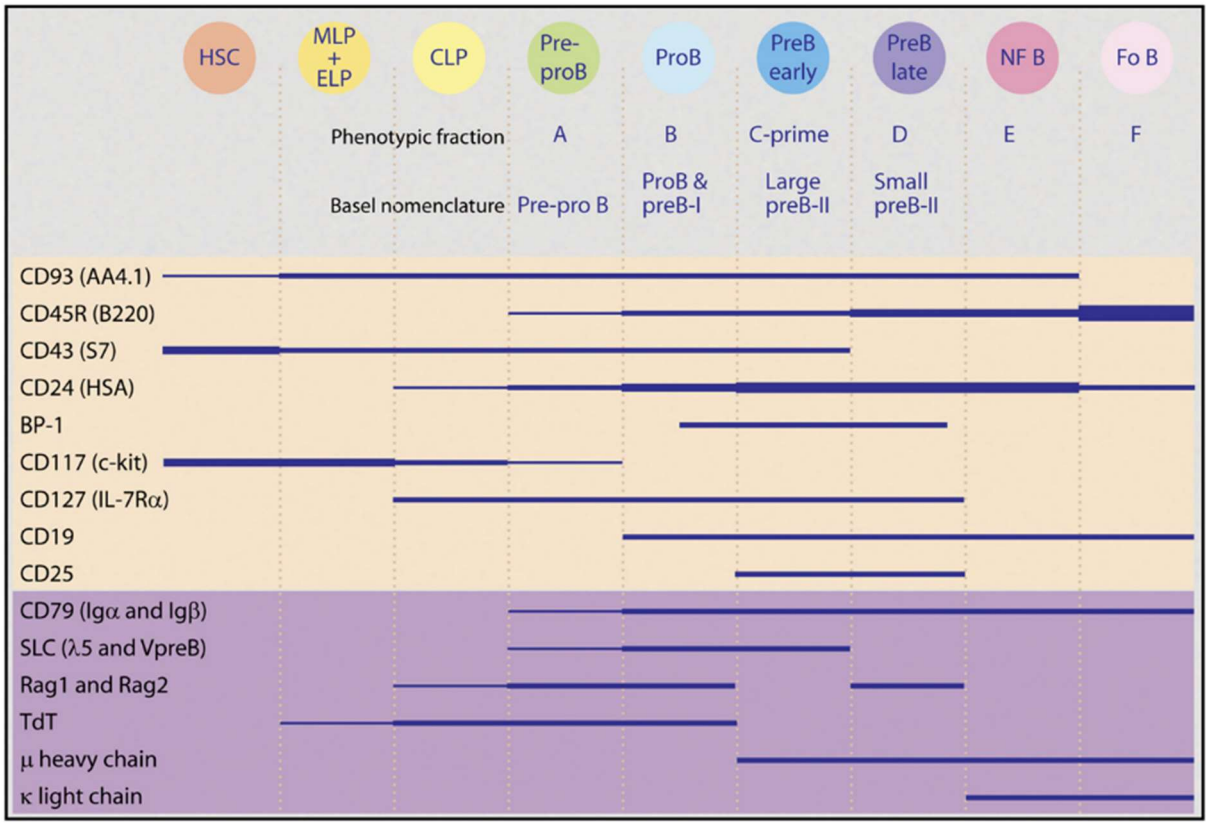


Figure 1-1: Changes in cell-surface molecules (beige) during murine bone marrow B cell development.
The stage-specific presence of different cluster of differentiation (CD) molecules has allowed their use as markers for separating B cell development into discrete stages. The changes in cell-surface markers are accompanied by changes in gene expression and intracellular protein translation (purple). Adapted from (Hardy et al., 2007).

1.2 Commitment to B cell development in bone marrow

1.2.1 VDJ recombination of the immunoglobulin heavy chain

After diversity (D_H) and joining (J_H) genes undergo a lineage specific recombination event (D_H-J_H) that was initiated in CLP, the sequential recombination of the variable (V_H) genes (V_H-D_HJ_H) follow in pro-B cells (fraction B/C) (Figure 1-2). This ordered lineage specific rearrangement of genes is known as V(D)J

recombination and is mediated by the RAG1/2 recombinases. RAG1/2 are responsible for the double-stranded DNA breaks at recombination signal sequence (RSS) sites that initiate recombination (Oettinger et al., 1990). The RSS is a genomic sequence composed of highly conserved heptamer (7 bp) and nonamer (9 bp) motifs that are separated by a 12 bp or a 23 bp spacer sequence (Figure 1-3 a). In the Igh locus, the V and J genes are flanked by the 23bp spacer while the D genes are flanked by the 12bp spacer. Recombination only efficiently happens between 12/23 RSSs (the '12/23 rule'), ensuring the ordered recombination of J to D to V genes and preventing the recombination of J genes to V genes (Schatz and Swanson, 2011). Recombination can take place by either deleting the intervening signal joint (SJ) sequence, in the case when the RSSs are in opposite direction, or the SJ sequence is inverted and leads to the retention of both the coding joint (CJ) and the SJ (Figure 1-3 b). The RSSs in the Igh all face opposite directions that results in the deletion of the SJ sequence, whereas half the Igκ V genes are in the same orientation as the $J_κ$ genes leading to recombination by inversion (Zachau, 1993) (Figure 1-3 b). The RSS, despite its conserved sequence, is almost never observed in its consensus configuration *in vivo*. Certain regions of the RSS are more important in recombination and display higher levels of conservation, such as the first three bases of the heptamer (Ramsden et al., 1994).

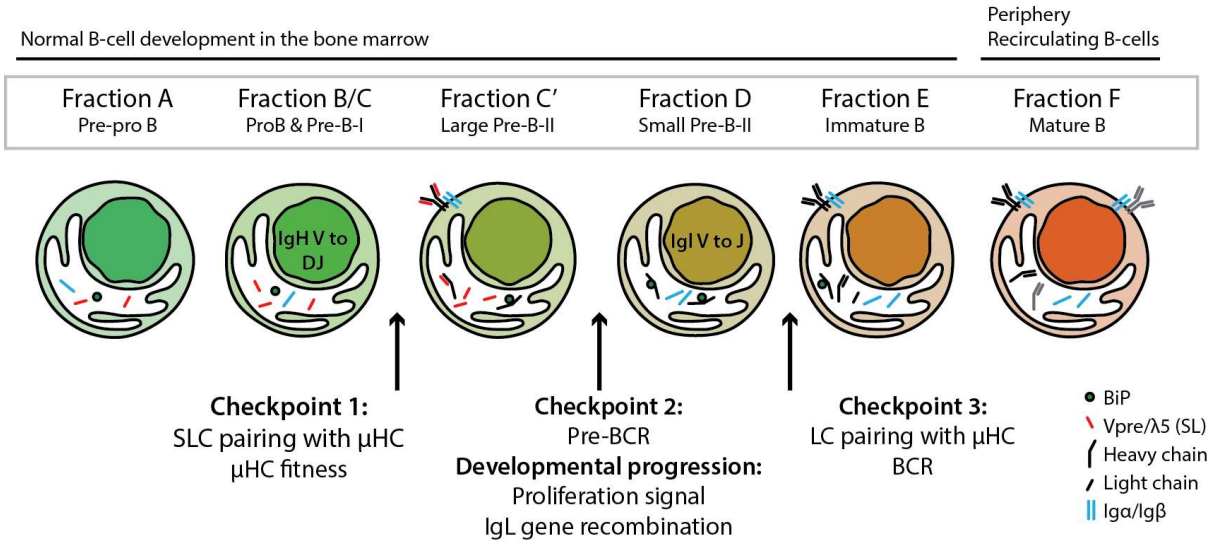


Figure 1-2: B cell development in the bone marrow.
The different stages of development as defined by the Philadelphia/Basel nomenclature and also described in terms of phenotypic fractions (Hardy et al., 1991) are shown along with key proteins, such as BiP and the surrogate light chain (SLC) components (VpreB and λ5), present at each stage. BiP is a chaperone protein that retains unpaired heavy chains within the endoplasmic reticulum (white). Igα/Igβ are signal transducers for the pre-BCR receptor. V to DJ recombination of the Igh terminates after fraction B/C, while recombination of the Igl takes places in fraction D. At three developmental stages (checkpoints) B cells are subjected to positive or negative selection. The first and second checkpoints relate to the pre-BCR receptor, during which its ability to assemble (checkpoint 1) and functionally signal (checkpoint 2) proliferation and recombination of the Igl is assessed. B cells without a functional pre-BCR and negatively selected out. The third checkpoint related to the BCR and the ability of the light chain to pair with the heavy chain. B cells with non-functional or autoreactive BCRs are selected out, while the rest are positively selected and allowed to mature in the periphery.

Beside the RSSs within immunoglobulin loci, there are also millions of other regions of the genome capable of binding the RAG1/2 recombinases termed cryptic RSSs (cRSS) (Schatz and Ji, 2011). cRSSs can, in cases such as in lymphoma, lead to chromosomal translocations. The exact sequence composition of the RSS has been proposed to be one of the main factors that determines recombination efficiency. To capture the potential of an RSS sequence for recombination, a statistical model was devised that assigns a recombination information content (RIC) score to each potential RSS. The RIC score was able to distinguish cRSSs from true RSSs as well as show a correlation with actively recombining V_H genes (Bolland et al., 2016; Cowell et al., 2002, 2003). However, other factors must also impact the process as observed from the different recombination frequencies of identical RSS sequences (Feeney et al., 2004). We have previously shown that the RSS acts more like a 'binary switch', enabling recombination of individual V_H genes, but other chromatin factors influence the efficiency/frequency of recombination (Bolland et al., 2016).

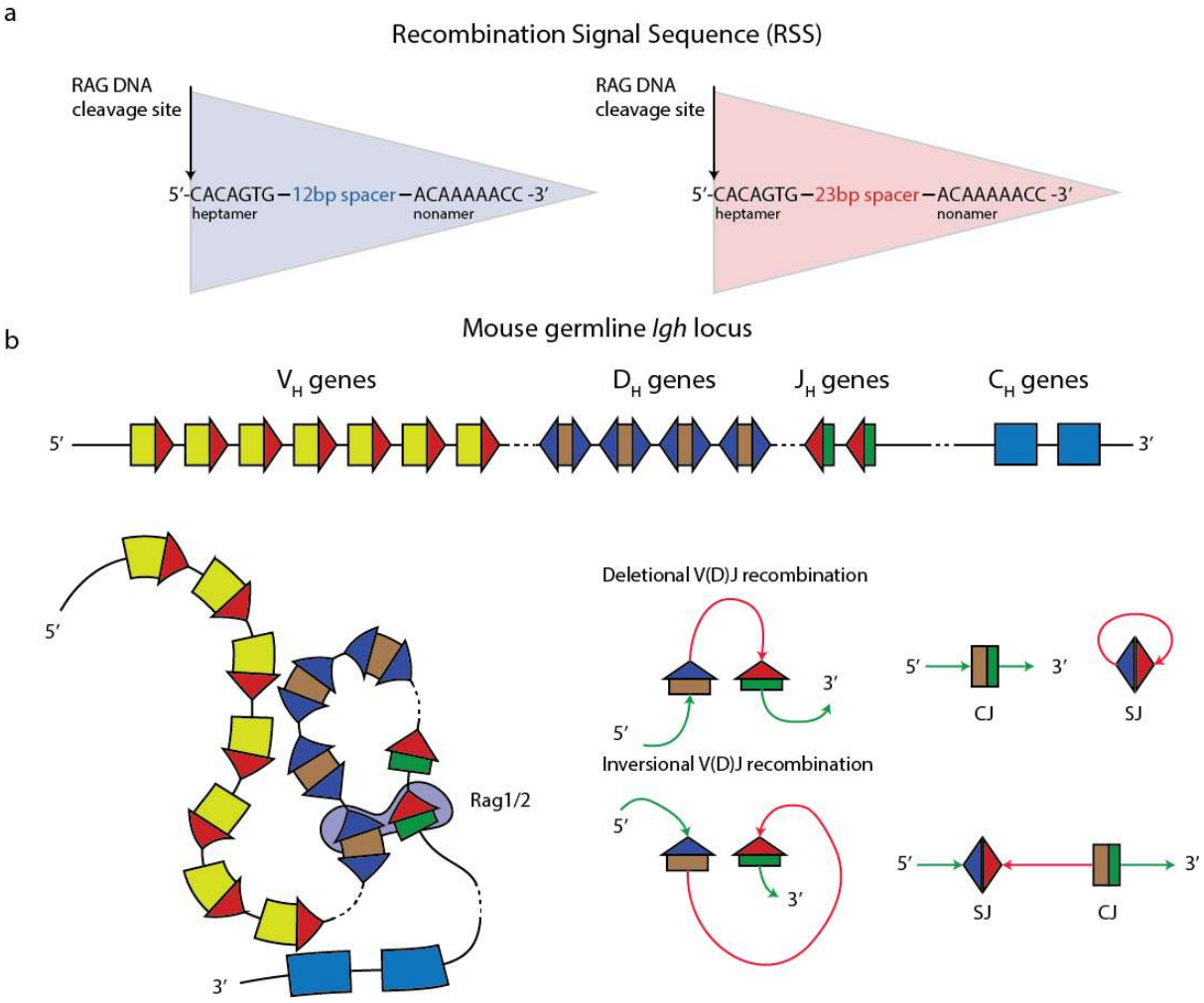


Figure 1-3: Recombination signal sequence (RSS) and RAG1/2 mediated recombination.
(a) The RSS contains a conserved heptamer and nanomer motif separated by 12 (blue) or a 23 bp (red) spacer. The difference between the spacers corresponds to one helical turn on the DNA. (b) During recombination, the RSS_{12} and RSS_{23} come together with a heterotetramer composed of RAG1 and RAG2. RAG1/2 cleaves the DNA sequence at the RSS and performs recombination through either deletion or inversion of the intervening DNA. In deletional recombination the signal joint (SJ)

sequence is circularised and removed, whereas in inversional recombination the SJ sequence is inverted and retained between the coding joint (CJ) sequence. Based on (Little et al., 2015).

Following the dsDNA cleavage by the RAG1/2 recombinases, a pair of hairpin loops form to protect the cleaved DNA. The hairpins are opened for repair with the Artemis:DNA-dependent protein kinase complex (Ma et al., 2002). If cleavage takes place at the apex of the hairpins, the result is a blunt-end. Alternatively, cleavage can take place several base pairs away from the apex, resulting in single-stranded overhangs. Filling of these overhangs results in palindromic DNA sequences (P nucleotides) that were not present in the germline. The cuts are subsequently repaired with the cooperation of the RAG recombinases with the non-homologous end joining (NHEJ) repair machinery along with TdT (Agrawal and Schatz, 1997; Schatz and Ji, 2011). The repair process leads to nucleotide deletions, which are characteristic of NHEJ, and TdT mediated non-template nucleotide additions (non-germline (N) nucleotide addition) (Desiderio et al., 1984). The addition of N and P nucleotides at the junctions between the immunoglobulin genes (V_H-Junction-D_H-Junction-J_H) further diversifies the antibody repertoire. The hyper-variability established through V(D)J recombination gives an antibody its specificity in the form of the complementarity determining region 3 (CDR3) (Tonegawa, 1983). Recombination is the first mechanism that drives the generation of the highly diverse antibody repertoire responsible for the recognition and neutralisation of pathogens.

1.2.2 Pre-BCR expression and signalling

The expression of the signal transducers Igα and Igβ along with the VpreB and λ5 non-covalently bonded proteins, which make up the surrogate light chain (SLC), happens early on in B cell development (Figure 1-2). After the productive recombination of the Igh, the resulting heavy chain (HC) polypeptide is deposited on the cell surface as a dimer along with Igα/β and the covalently bonded SLC, forming the pre-BCR. After numerous rounds of B cell receptor selection during development, only a minority of B cells from the starting pool manage to reach maturity (Allman et al., 1993). This selection takes place at several key developmental checkpoints.

1.2.2.1 Surrogate light chain pairing and the pre-BCR - checkpoint 1 and 2

The pre-BCR constitutes the first major selection checkpoint during B cell development (Figure 1-2). The pairing of the SLC with the HC indirectly assesses the ability of the HC to ultimately pair with the light chain (LC) peptide at the immature B cell stage (fraction E) and form a functional BCR receptor (Melchers, 1999). The expression of the pre-BCR triggers a signalling cascade that leads to proliferation and developmental progression. CD25, the α chain of the interleukin-2 receptor (IL-2Rα), is transiently upregulated on the cell surface of pre-BCR expressing B cells (Figure 1-1) along with CD2, while RAG1 and RAG2 are down-regulated.

The functional relevance of CD25 upregulation on pre-B cells is unknown. Culturing B cells in the presence of IL-2 does not affect their proliferation or differentiation (Rolink et al., 1994), and a three cell surface marker analysis of bone marrow B cells in IL-2Rα knockout mice showed no differences from their wildtype counterparts in young mice (Willerford et al., 1995). With age, both IL-2 deficient and IL-2Rα knockout mice display a severe reduction of B cells in the mature compartments and throughout early development (Schultz et al., 2001; Willerford et al., 1995). The loss of both IL-2Rα positive B cells and B cells that do not yet express IL-2Rα suggests an IL-2Rα independent cause, one which has been suggested to involve an influx of dysregulated T cells into the bone marrow (Sadlack et al., 1995; Schultz et al., 2001). An abstract published from an ASH annual meeting suggest that the two other chains that make up the IL-2 receptor, IL-2Rβ (CD122) and IL-2Rγ (CD132), do not pair with IL-2Rα and do not respond to the IL-2 ligand. However, it additionally suggests that IL-2Rα is involved in pre-BCR signalling, revealing an unappreciated role of this receptor in B cell development once the data is published (Lee et al., 2015).

Large pre-B cells (fraction C') undergo 2-5 cell divisions (clonal expansion) during which the absence of the recombinase enzymes is critical (Decker et al., 1991; Hess et al., 2001). This expansion is essential for increasing the immunoglobulin repertoire diversity, as multiple HC clones can each have a differently rearranged LC. The expansion and differentiation has been shown to be independent of the pre-B cell environment, which suggests it can be driven by internal signalling events (Rolink et al., 2000). The pre-BCR also signals its own expression-termination and degradation by inducing the expression of the lysosome-associated protein transmembrane 5 (LAPTM5). LAPTM5 was shown to target a large pool of pre-BCRs within the endoplasmic reticulum (ER) to the lysosome for degradation (Parker et al., 2005; Kawano et al., 2012). The degradation along with proliferative dilution leads to the observed absence of the pre-BCR in small pre-B cells (fraction D) (Figure 1-2).

1.2.2.2 Pre-BCR signalling regulation

SLP-65 (also known as BLNK/BASH) is the main signalling molecule downstream of the pre-BCR that mediates the SLC negative feedback loop, opposes the termination of heavy chain recombination and inhibits proliferation induced by class IA PI3K signalling (Herzog et al., 2009) (Figure 1-4) to essentially allow the recombination of the light chain. The pre-BCR feedback loop involves the up-regulation of Ikaros, Aiolos and IRF4 via SLP-65, which subsequently silence SLC genes (Ma et al., 2008; Thompson et al., 2007) (Figure 1-4). SLP-65 expression is crucial in controlling proliferation, developmental progression and B cell function, as highlighted by SLP-65 knockout mice (Jumaa et al., 1999; Hayashi et al., 2000). The SLC greatly increases the signal transduction of the pre-BCR via SLP-65 (Meixlsperger et al., 2007) and in turn silences itself through the downstream Ikaros and Aiolos transcription factors. The exact mechanism by which SLP-65 represses PI3K is not yet known. The phased timing of large

pre-B cells (fraction C') proliferation induced by PI3K, before its repression by SLP-65, has been suggested to be the result of the abundance of PI3K proteins at the time of pre-BCR signalling versus the need for expression and translation of SLP-65 that delays its presence (Herzog et al., 2009). When the pre-BCR initiates signalling, it leads to the phosphorylation of SYK (autonomous receptor signalling), which in turn activates the PI3K pathways that results in the degradation of forkhead box O (FOXO) transcription factor (TF) in the absence of SLP-65. Absence of FOXO TFs leads to proliferation and inhibition of differentiation, whereas its presence leads to RAG1/2 expression, Igl recombination and B cell differentiation (Herzog et al., 2009). SLP-65 is crucial in differentiation and stopping proliferation, acting as a tumour suppressor. The cell cycle is controlled by cyclin-dependent kinases (CDKs) that are dependent on cyclins to become catalytically activated. Cyclin D3 has been shown to be involved in the expansion of the large pre-B cell pool (fraction D) in a receptor-mediated manner. The pre-BCR is required to stabilise cyclin D3 and prevent its degradation (Cooper et al., 2006). FOXO has been implicated in cell cycle delay through the regulation of p27 (a CDK inhibitor) expression. In addition, FOXO activates the expression of RAG1/2 during p27 induced cell cycle arrest (Herzog et al., 2009). In the absence of p27, RAG2 is targeted for degradation by the cyclin A/CDK2 complex (Lee and Desiderio, 1999) (Figure 1-4). Through this mechanism, recombination is confined to G_1 and G_0 of the cell cycle, circumventing risks associated with replication.

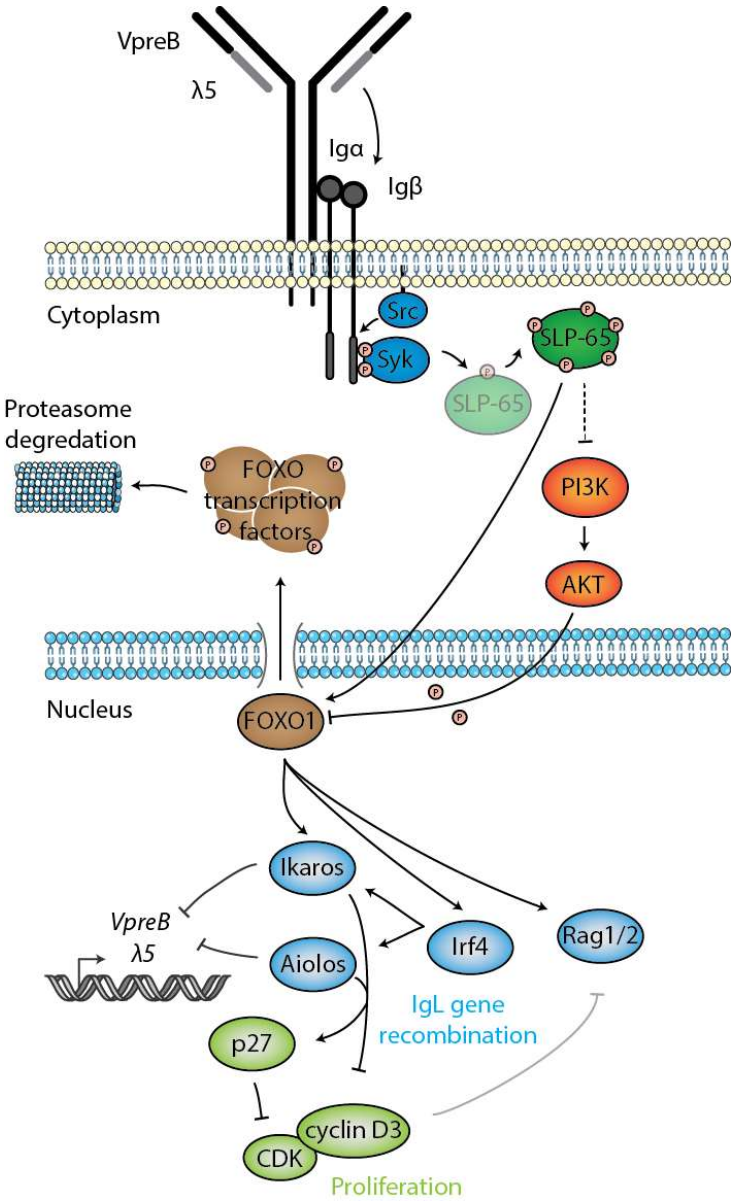


Figure 1-4: Pre-BCR signalling activation of SLP-65.
Syk dependent phosphorylation and activation of SLP-65 leads to the repression of PI3K activity through a yet unknown mechanism and the activation of the FOXO1 transcription factor. In the absence of SLP-65, PI3K through AKT (protein kinase B or PKB) phosphorylates the FOXO transcription factors that results in their relocalisation into the cytoplasm and degradation by the proteasome. After the proliferative burst supplied through the pre-BCR and mediated by PI3K, SLP-65 assumes control and turns SLC (VpreB, λ5) expression off, leading to the downregulation of surface pre-BCR. SLP-65 also initiates IgL recombination through FOXO1 mediated cell cycle arrest, which involves the inhibition of cyclin D3 via Ikaros and Aiolos, and the upregulation of Rag1/2. Aiolos also promotes expression of the CDK inhibitor p27. CDK signal the cell that it can progress into the next stage of the cell cycle. The inhibition of the CDK-cyclin complexes also prevents the degradation of Rag2, which is normally ubiquitylated and targeted for degradation via the CDK-cyclin complexes (Herzog et al., 2009). Based on (Werner and Jumaa, 2015).

1.2.2.3 The role of polyreactivity in pre-BCR function and BCR editing

The importance of polyreactive BCR specificity for B cell development was demonstrated in a study by Köhler *et. al.* using an inducible system of the downstream target of the pre-BCR and BCR receptor, SLP-65 (Figure 1-4). By expressing a fusion of a polyreactive or a non-autoreactive V region with the μHC and inducing SLP-65 in a SLP-65$^{-/-}$ line, they were able measure the ability of the polyreactive

chimeric receptor to induce calcium flux. The non-autoreactive receptor on the other hand, was only able to induce efficient calcium mobilisation after stimulation with an anti-µHC antibody (Köhler et al., 2008). The ligand-independent signalling ability of polyreactive BCRs even in the presence of the LC highlighted their similarity with the pre-BCR, for which the SLC has been shown to strongly enhance the induction of calcium flux, while the LC reduces it (Meixlsperger et al., 2007). In addition, the study showed that with artificial receptor editing (replacing the LC) the autonomous signalling of polyreactive BCRs was abolished. Overall, the study suggests that auto/polyreactivity plays a key role in the function of the pre-BCR and additionally facilitates the expansion of early immature B cells (fraction E) (Köhler et al., 2008).

1.2.3 The BiP chaperone predominantly bind the C_H1 antibody domain

Conventional unpaired µHCs are efficiently retained within the endoplasmic reticulum (ER) through one of the Hsp70 chaperone family (heat shock) proteins, called BiP, that is present at high concentrations within the ER (Haas and Wabl, 1983; Ellgaard and Helenius, 2003). BiP binds the C_H1 domain of the heavy chain (Bole et al., 1986; Hendershot et al., 1987) and, unlike other antibody domains that can fold independently of each other, has been characterised *in vitro* as intrinsically disordered (Feige et al., 2009). The HCs are retained within the ER (Hendershot et al., 1987) in a reduced state until the constant region of the light chain releases BiP in an ATP dependent manner (Vanhove et al., 2001) and folds the domain into its correct conformation (Feige et al., 2009). Although BiP predominantly interacts with the C_H1 domain, the deletion or replacement of the domain demonstrated that BiP continued to weakly bind the altered HC (Kaloff and Haas, 1995). Indeed, in some cases of heavy chain disease (HCD), where only the V region of the HC is missing, the ability to express the HCs on the cell surface was demonstrated (Corcos et al., 1991, 1995). It is the additional formation of dimers between HCs lacking the C_H1 domain that leads to the release of residual BiP (Kaloff and Haas, 1995). A similar stabilisation of the C_H1 domain in HCs from HCD patients that lack the V region could explain their ability to be expressed on the cell surface (Corcos, 2010). Interestingly, HC without the C_H1 domain were shown to still bind the LC at the C_H2 and C_H3 domains non-covalently (Hendershot et al., 1987). Overall, this illustrates the importance of the C_H1 domain, together with the SLC or the LC, as a safeguard to ensure ordered progression that is essential in normal B cell development. In a LC knockout and HCD models, B cell development fails past the immature B cell stage (fraction E) (Corcos et al., 1995; Zou et al., 2003). Interestingly, the ablation of the SLC does not fully block development at the pro-B (fraction BC) stage (Mundt et al., 2001a; Schuh et al., 2003; Shimizu et al., 2002; Su et al., 2003) despite the expectation that the HC should be sequestered in the ER via BiP (Shimizu et al., 2002).

1.2.4 The BCR and central tolerance - checkpoint 3

The re-expression of the recombination machinery (RAG1/2) after the termination of pre-BCR signalling initiates V_L-J_L recombination of the Igκ (kappa light chain) locus in small pre-B cells (fraction D) (Oettinger et al., 1990) (Figure 1-2) and results in the formation of the BCR receptor. The enormous diversity generated by VDJ recombination inevitably results in a subset of immunoglobulins with self-recognition. High levels of signalling through the BCR receptor leads to the activation of proliferation instead of differentiation at the immature B cell stage (fraction E). This seems to be a mechanism to allow further Igl recombination with the aim of reducing the high avidity of auto-reactive BCRs for self-antigens (Pelanda et al., 1997; Meixlsperger et al., 2007). In such cases, the recombination machinery that was retained from the initial Igl recombination is used for successive rounds of Igκ editing or recombination of the second Igl locus, Igλ (lambda light chain). Receptor editing is one of the main mechanism of central tolerance, along with cell inactivation (anergy) and the last resort of clonal deletion (apoptosis) (Nemazee, 2006; Pelanda and Torres, 2012). In addition to the negative selection of high level signalling BCRs, the absence of tonic signalling from unproductive BCRs has been shown to trigger receptor editing and lead to a de-differentiation phenotype (Schram et al., 2008; Tze et al., 2005). This additional trigger allows the rescue of B cells that contain a LC incapable of pairing with the HC. Tonic signalling would normally silence recombination and promote positive selection (Verkoczy et al., 2007) and its ablation ultimately results in clonal deletion (Lam et al., 1997). Successful surface expression of a functional BCR, composed of two HCs covalently paired with two LCs, terminates RAG1/2 expression and gives rise to the immature B cell (fraction E) that can migrate out of the BM and further mature in the periphery (Melchers, 2015).

1.3 Chromatin states governing VDJ recombination in the *Igh locus*

The expression of the RAG recombinase enzymes is only observed in specific lineages of lymphocytes and at precise developmental stages. An additional layer of control atop RAG expression is apparent from the sequential recombination of D_HJ_H before V_HDJ_H. The chromatin states that exert temporal control over recombination through chromatin accessibility will be discussed in this section, with a focus on the mouse Igh locus. Not much is known about the chromatin states in the human context, knowledge of which will be crucial in extending our understanding of how VDJ recombination is controlled in humans and whether it mirrors the mechanisms uncovered in mouse.

1.3.1 Igh locus relocation and regulation during development

The lamina network is essential to the architecture and function of the nucleus. Lamina-binding proteins serve as an interface between the nuclear backbone and its contents. They have been implicated in influencing a range of functions from transcriptional activity, epigenetic marks to location (Wilson and Foisner, 2010). Constitutive heterochromatin and the nuclear periphery are two

major classes of transcriptionally repressed sub-nuclear compartments. They are characterised by condensed chromatin and play an important role in chromatin accessibility and gene suppression (Pickersgill et al., 2006; Beisel and Paro, 2011). Previously, the importance of enhancers in locus relocation has been demonstrated with the β-globin locus (Ragoczy et al., 2006). The Igh locus contains a super-enhancer, the intronic enhancer Eμ, which was suspected to have a role in relocation (Bowen and Corcoran, 2008). In non-lymphocytes, the immunoglobulin loci are sequestered at the nuclear periphery. The dependence of chromosomal relocation, from the nuclear periphery into the inner nucleus, on the intronic enhancer Eμ was later shown using 4C and related techniques (Guo et al., 2011a). 4C is a proximity ligation based assay that captures chromosome conformations in a 'one versus all' format, meaning it captures all the fragments interacting with a site of interest (Wit and Laat, 2012). The most distal part of the murine Igh locus (5' J558 genes) is localised at the nuclear periphery, while the 3' end protrudes into the inner nucleus (Yang et al., 2005). This 3' DJ protrusion is thought to be recruited to a transcription factory, where the continuous transcription of Iμ from the intronic enhancer Eμ anchors it in place (Corcoran, 2010). The term transcription factory stems from the observation that transcription of active genes occurs in pre-formed nuclear foci with high concentrations of active RNA polymerase II (RNAPII) (Hozák et al., 1993; Mitchell and Fraser, 2008). The localisation of the Igh at a transcription factory may contribute to ordered recombination of the VDJ genes (Bowen and Corcoran, 2008).

1.3.2 Allelic exclusion

Each B cell contains a unique immunoglobulin that is selected to recognise and bind to a single antigen with high affinity. This one cell equals one receptor nature of B cells is achieved through a process called allelic exclusion, which collectively describes the events that lead to the productive rearrangement and expression of only a single uniquely recombined Igh and Igl allele within each B cell. Allelic exclusion involves the sensing of a successful recombination through multiple mechanisms, including DNA damage-sensing proteins, RAG2, mRNA and signalling triggered by the pre-BCR.

Unlike D_H-J_H, V_H-D_HJ_H recombination is limited to only one allele despite both alleles being in euchromatic regions of the nucleus and accessible to the recombination machinery. The monoallelic restriction of V_H-D_HJ_H recombination has been attributed to the recruitment of ATM (ataxia-telangiectasia mutated) to the site of the RAG-mediated dsDNA break, triggering signalling that leads to the repositioning of the second allele to pericentromeric heterochromatin (PCH) (Hewitt et al., 2009). Additional support for the role of allele repositioning in allelic exclusion was provided by the same lab in a study where the absence of ATM or the C-terminus of RAG2 resulted in a higher frequency of biallelic breaks and the retention of both alleles in the euchromatic compartment (Chaumeil et al., 2013). The exact mechanism by which ATM and RAG2 perform this relocalisation is

unknown, but allele-specific 4C experiments in mature B cells showed both alleles present within similarly active compartments (Holwerda et al., 2013), suggesting the PCH relocation is only transient. Interestingly, in mature B cells the distal V region of both alleles appears to be recruited to a less transcriptionally active compartment, which has been proposed to focus the transcriptional and enhancer activity onto the productively recombined V promoter (Holwerda et al., 2013).

The triggering of allelic exclusion by sensing dsDNA breaks does not explain the ability of B cells to recombine the second allele in cases of unproductive recombination. A study that separated the transcription of the recombined immunoglobulin gene from the protein production was able to show that a μHC transgene placed within one of the endogenous alleles and under the control of a weak promoter induces recombination of the second allele. In the case of a strong promoter, secondary recombination was inhibited (Lutz et al., 2011). The transcription of an unproductive recombination results in the rapid degradation of the transcript by the nonsense mediated decay pathway (Eberle et al., 2009). The study by Lutz et al. suggests that the stability of transcripts and therefore the abundance of mRNA allows cells to distinguish productive from unproductive recombination events that determine if further recombination is required or if allelic exclusion can get underway.

After the relocalisation of the Igh into a transcription factory, it undergoes compaction that is dependent on interleukin-7 receptor (IL-7R) signalling (Kosak et al., 2002). The compaction is required to equalise the probability of recombination of distal V_H genes, which are located up to 2.5 Mb away from the D region proximal V_H genes. The subsequent productive recombination and formation of the pre-BCR triggers the degradation of RAG1/2 (Figure 1-4), until the recombination of Igl, along with a reduction of IL-7 signalling that reduces the accessibility of the V_H region (Chowdhury and Sen, 2003). At the same time, both Igκ alleles undergo decompaction, mediated by the 3' Igκ enhancer, that suggests a mechanism for limiting downstream recombination (Hewitt et al., 2008; Roldán et al., 2005). The weakening of IL-7 signalling allows the activation of the Igκ intronic enhancer (iEκ) through the increased binding of the E2A transcription factor (Johnson et al., 2008). The dependence of Igh recombination on IL-7 is contrasted by its negative regulation of Igκ recombination, highlighting a mechanism that ensured a sequential non-overlapping recombination of the heavy and light chain loci during the transition between the pro- (fraction BC) to pre-B (fraction C'D) developmental stages. Altogether, an exquisite interplay of numerous mechanisms contributes to the process of allelic exclusion.

1.3.3 Non-coding RNA transcription

Protein coding transcription constitutes only a fraction of the entire mammalian transcriptome. The importance of non-coding RNAs (ncRNA) in transcriptional regulation is becoming more apparent as our understanding of the role of ncRNAs expands. Their multitude of functions is mirrored by their

diversity. Some of the main categories of ncRNAs include miRNA, siRNA and long non-coding RNAs (lncRNA) (Mercer et al., 2009). The Igh locus contains extensive lncRNA transcription. The Iμ transcripts that originate from the intronic enhancer Eμ (Lennon and Perry, 1985) are among the first transcribed in the Igh loci, along with the *μ0* transcripts, just prior to recombination. *μ0* emanates from the PDQ52 promoter/enhancer located upstream of the first D_H gene (DQ52) (Kottmann et al., 1994) (Figure 1-5).

1.3.4 Antisense intergenic transcription and histone modification

Intergenic transcription has been identified as a key component in facilitating VDJ recombination. The Igh locus contains both V region and D region antisense intergenic transcription (Figure 1-5) that has been shown to precede recombination and has been proposed to remodel the chromatin to enable recombination (Bolland et al., 2004, 2007). Through the activity of the RNA polymerase II complex, the locus is made accessible to chromatin re-modellers. The initiation of transcription in closed chromatin has been shown to be dependent on the RNA polymerase II (RNAPII) associated SWI/SNF ATP-dependent remodelling complex (Li et al., 2007; Osipovich et al., 2009). Opening of chromatin facilitates other changes that include loss of the repressive histone H3K9 methylation and gain of permissive H3K36me3, H3 and H4 acetylation marks (Subrahmanyam and Sen, 2010). Antisense transcription is present in discrete locations within the Igh locus and correlates with binding sites of regulatory factors involved in looping (Choi et al., 2013a). A model has been proposed where long antisense intergenic transcription shares a transcription factory with Eμ and in so doing brings the distal V_H region in close proximity of the DJ_H region (Stubbington and Corcoran, 2013).

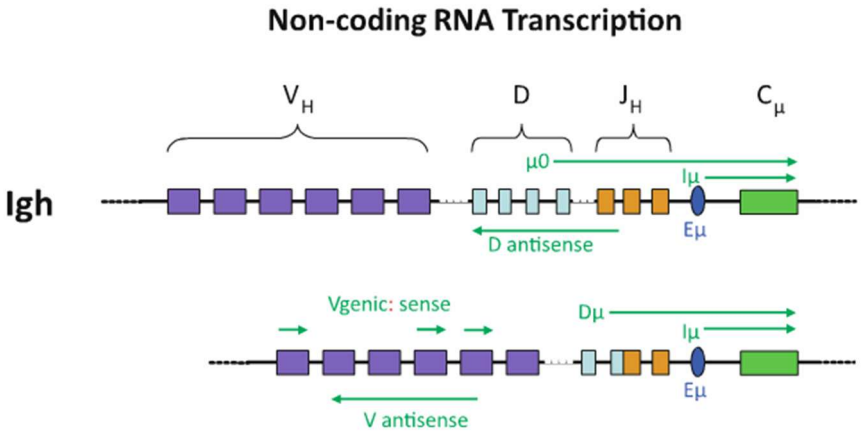


Figure 1-5: Non-coding transcription of the Igh locus sequentially activated during VDJ recombination.
Top figure depicts the germline Igh with transcripts in green and the Eμ enhancer depicted as purple oval. Bottom figure depicts a D_HJ_H recombined Igh locus with transcription initiated in the V_H region. Taken from (Matheson and Corcoran, 2012).

1.3.5 Igh regulatory elements

The Igh has several enhancers and insulators that regulate the diverse chromatin landscape. Some of these include the Eμ enhancer and the PDQ52 promoter/enhancer, as mentioned above. The Eμ enhancer is indispensable in VDJ recombination. Besides its role in relocation and transcription factory tethering, its targeted deletions results in impaired D_H to J_H recombination (Afshar et al., 2006) and ablation of D_H antisense transcription (Bolland et al., 2007). Unlike Eμ, the deletion of PDQ52 only results in modest alteration of J gene usage and has been proposed to have importance in secondary recombination (Nitschke et al., 2001). In addition to these two enhancers, a 3' regulatory region (3'RR) composed of four DNase I hypersensitive site (hs1-4) has been implicated in the regulation of numerous processes. Most prominently, the 3' RR region is involved in class switch recombination (CSR), promoting high transcription output of plasma cells and providing accessibility to the somatic hypermutation machinery. Upon deletion of the entire region all three functions of the 3'RR are severely compromised (Rouaud et al., 2013; Vincent-Fabert et al., 2010). In addition to the hs1-4 sites, there are four additional hypersensitive sites (hs5-8) downstream that are interspersed with a high density of CTCF (CCCTC-binding factor) and PAX5 binding sites (Birshtein, 2012). CTCF has been implicated in numerous roles ranging from transcriptional activation, repression, insulation to genome architecture (Phillips and Corces, 2009). The other regulatory region identified in the Igh locus is the IGCR1 insulator composed of two CTCF binding sites: HS4 and HS5. It is located in the V_H-D_H intergenic region and was shown to substantially decrease antisense transcription in the V_H region (Featherstone et al., 2010). Chromatin conformation capture (3C; 4C; Hi-C) studies of the Igh revealed multiple interactions between the 3'RR and the rest of the locus (Birshtein, 2014), with a notable interaction being formed between the hs8 site and the IGCR1 (Medvedovic et al., 2013). In addition, the 3'RR has been shown to overlap the boundary of a topologically associated domain (TAD) that insulated the locus from downstream genes (Benner et al., 2015). The IGCR1 has been shown to inhibit proximal and promotes distal V_H gene rearrangement. Perhaps most importantly it is responsible for maintaining a lineage specific and ordered recombination of D_HJ_H followed by V_H-DJ_H recombination (Guo et al., 2011b). The knockout of the hs5-7 sites, leaving hs8 behind, was shown to produce a mild skew in the normal choice of recombining genes (Volpi et al., 2012), suggesting that the overall interactions involving the 3'RR may influence VDJ recombination.

Two regulatory regions equivalent to the mouse 3'RR were also discovered in the human IGH locus downstream of each constant α gene. They contain three DNaseI hypersensitive sites that are equivalent to the mouse hs1-4, with the exception of the mouse hs3B that seems to be absent in humans (Mills et al., 1997). Besides the 3'RR, the human IGH also shares the Eμ intronic enhancer with the mouse Igh (Hayday et al., 1984; Staudt and Lenardo, 1991). In addition, a difficult to clone region

between the Cδ and Cγ3 genes was shown to contain a high density of transcription factor binding motifs separated by repeats. A luciferase reporter assay revealed potential activating and silencing activity of the sequence, while an *in vivo* transgene revealed its activity was constrained to bone marrow B cells, suggesting it may represent a human specific enhancer (Mundt et al., 2001b).

1.4 Evolution of the Igh locus

The murine Igh locus of C57BL/6 mice is composed of 195 V_H (110 functional, 85 pseudogenes), 10 D_H, 4 J_H and 8 constant (C_H) region genes that overall span approximately 3 Mb (Johnston et al., 2006; Ye, 2004). Out of the 3 Mb, approximately 2.5 Mb is occupied by V_H genes that can be divided into 15 families based on their sequence homology (Mainville et al., 1996). In contrast, the human IGH locus only spans approximately 1.2 Mb and contains 123 V_H genes (44 functional, 79 pseudogenes), 27 D_H genes (23 functional, 4 pseudogenes), 9 J_H genes (6 functional, 3 pseudogenes) and 5-11 constant region genes depending on haplotype (Lefranc and Lefranc, 2001; Matsuda et al., 1998). The human V_H genes can be divided into 7 families (Kodaira et al., 1986). The Igh gene composition within different species arose from species-specific duplication events and subsequent divergence. Duplicated genes can either acquire a new function and remain in the genome or become pseudogenes as proposed in the birth-and-death evolution model (Ota and Nei, 1994). Comparison of sequence similarity in the C57BL/6 mouse strain showed that V_H genes can be grouped into three distinct phylogenetic clans (Johnston et al., 2006). This confirmed the previous assignment of the mammalian V_H families into three clans based on framework 1 (FR1) protein sequence homology (Kirkham et al., 1992). Despite the obvious size and gene number difference between human and mouse, it is remarkable that their V_H genes can be assigned to one of the three shared clans that have survived in the genome for millions of years (Das et al., 2008; Ota and Nei, 1994) (Figure 1-6).

Figure 1-6: Distribution of the three clans across the V_H genes in human and mouse.
In green are the D_H genes. Red – clan I, blue – clan II, yellow – clan III. Adapted from (Das et al., 2008).

1.4.1 Chromatin states driving recombination

Numerous VJ primer pair PCR based methods have been used to study antigen receptor (AgR) repertoires, most recently with adaptations for next-generation sequencing (NGS) (Georgiou et al., 2014; Weinstein et al., 2009). However, the biases present within these techniques hinder the quantification of the antibody repertoire diversity (Baum et al., 2012). The development of a novel assay termed VDJ-seq allowed us to overcome these limitations. VDJ-seq uses the fact that in every recombination there will always be a J_H gene present and leverages this to perform a single round of

primer extension that extends into the D_H and V_H recombined sequence. The primer extension is performed with biotinylated primers that are captured, amplified and made into an Illumina ready library (Bolland et al., 2016; Chovanec et al., 2018).

Obtaining quantitative data of DJ_H and VDJ_H recombination frequencies in mouse, allowed us to perform a detailed examination of the efficiency of this process. ChIP-seq datasets of multiple transcription factors and histone modifications along with other chromatin data, such as DNase hypersensitivity and non-coding RNA transcription, were used to determine the chromatin state of each V_H gene. By integrating sites of active recombination within the Igh together with the chromatin state annotation from ChromHMM (Ernst and Kellis, 2012), it was possible to start unravelling the states of chromatin that influence recombination in the mouse Igh. The analysis resulted in a model that separated V_H genes into three mutually exclusive chromatin states (State A: CTCF, RAD21; State E: PAX5, IRF4, YY1, along with various histone modifications; Background state). The model was highly predictive of recombination. Interestingly, the mutually exclusive A and E chromatin states correlated with the evolutionary history of the locus (Bolland et al., 2016). An equivalent analysis was also performed for the Igκ locus revealing that PU.1, not important in the Igh, acts as a binary predictor of recombination in the Igκ locus (Matheson et al., 2017).

1.5 B cell maturation in the spleen

Immature B cells produced in the BM migrate into the spleen where they either die or undergo maturation. Studies with transgenic mice have shown a dependence on BCR signalling for the transition into different mature subtypes. Specific signalling (BAFF or Notch2) pathways, including the BCR signal strength, seems to be the driving force behind lineage progression into the different mature B cell populations (FO, section 1.5.2; MZ, section 1.5.3; B1, section 1.5.4) (Hardy et al., 2007) (Figure 1-7).

1.5.1 T1-3 B cells

Much like the development in the bone marrow, B cell maturation in the spleen has been shown to be a multistep process with a number of selection checkpoints (Allman et al., 2001; Loder et al., 1999). The progression from immature to mature B cells can be subdivided, based on differential expression of cell-surface markers, into three 'transitional' stages: T1, T2 and T3. They are all short-lived cells and still functionally incompetent as they are unable to respond to antigen stimulation, which is a prerequisite of mature B cells (Hardy et al., 2007).

1.5.1.1 T3 Anergy - checkpoint 4 (peripheral tolerance)

The immature B cells entering the spleen have been shown to undergo ligand-dependent apoptosis (Rolink et al., 1998) and silencing by anergy (Merrell et al., 2006). The T3 stage is characterised by down-regulation of IgM and was thought to have represented an intermediate stage in the generating

follicular B cells (Allman et al., 2001). It was later shown that the T3 stage contains functionally unresponsive (anergic) B cells, based on a panel of cell surface markers of anergy ($CD93^{+}$, $CD24^{intermediate}$, IgD^{high}, $CD23^{+}$, IgM^{low}) that was defined from a comparative analysis of wildtype and anergic transgenic mouse strains (Merrell et al., 2006). Therefore, BCRs with low affinity for auto-antigens passing through the tolerance checkpoint during the transition from immature to mature B cells are thought to be diverted into this anergic state. Continuous presence of auto-antigens is required to keep them in this state (Merrell et al., 2006). This represents an additional checkpoint in the periphery for selecting out auto-reactive BCRs, which seems to be more prominent in humans than in mice (Melchers, 2015).

1.5.2 B2 Follicular (FO) B cells

The final stage of maturation of a large population of recirculating immature B cells takes place in the anatomical site called the follicular region. Cells within this region are referred to as follicular B cells. They do not proliferate but constitute a population of long-lived B cells. Their survival is dependent on B cell-activating factor (BAFF) signalling more so than on BCR signalling (Figure 1-7) (Hardy et al., 2007). FO B cells are thought to be the main mature B cells population that give rise to long lived memory B cells (Shlomchik and Weisel, 2012; Victora and Nussenzweig, 2012).

1.5.2.1 Somatic hypermutation (SHM) - checkpoint 4 (peripheral tolerance)

Antigen stimulation of FO B cells leads to proliferation and the expression of activation-induced cytidine deaminase (AID). AID catalyses the switching of antibody effector class (class switch recombination) and the honing of antibody binding avidity through SHM. The process of SHM may lead to auto-reactive antibodies which need to be silenced and removed from the repertoire (Melchers, 2015).

1.5.3 Marginal zone (MZ) B cells

MZ B cells are another population of mature B cells that are located in areas occupied by macrophages, which classically perform antigen filtering and scavenging tasks. Much like B-1 B cells (section 1.5.4), MZ B cells have been suggested to have a restricted repertoire (Martin and Kearney, 2002; Pillai et al., 2005). Unlike FO B cells, MZ B cells are specialised against bacteria cell-wall constituents and senescent self-components (Hardy et al., 2007; Paul, 2012). The clearance of apoptotic cells and some types of bacteria is thought to be aided by their low level of polyreactive and autoreactive specificity (Martin and Kearney, 2002). Notch2 signalling is indispensable in MZ B cells as its deletion leads to MZ B cell absence (Saito et al., 2003), contrasting FO B cells dependence on BAFF (Figure 1-7).

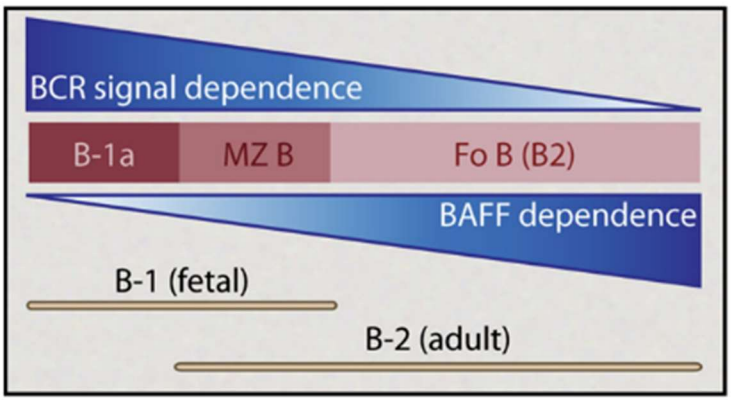


Figure 1-7: The reciprocal signal dependence of different mature B cell subtypes. Taken from (Hardy et al., 2007).

1.5.4 B-1 B cells

B-1 B cells seem to be exclusively generated by the fetal liver, as demonstrated by transplant experiments (Hayakawa et al., 1985). They persist primarily through self-renewal, unlike FO and MZ B cells that are derived from BM precursors (Hayakawa et al., 1986). They produce what has been termed natural autoantibodies, as they are found even in mice kept in sterile conditions. Natural antibodies have higher specificity for self, compared to antibodies produced by B2 B cells (Hardy et al., 2007). Autoreactive specificity seems to be favoured during selection, allowing self-reactive BCRs into the mature B-1 B cell pool (Baumgarth, 2011).

1.6 Therapeutic antibodies

The use of monoclonal antibodies (mAbs) as therapeutics began in the 1980s and has grown into a multi-billion dollar industry with over 40 monoclonal products having gained clinical approval (Ecker et al., 2015). The pioneering hybridoma technology that allowed the isolation of mAbs from mice (Köhler and Milstein, 1975), along with the relative ease of mouse immunisation, resulted in extensive research and utilisation of rodent monoclonal antibodies as therapeutics. However, even before the advent of mAbs, serum therapies that used the sera from immunised non-human sources resulted in adverse effects termed serum sickness (Yamada, 2011). Similarly, mouse mAbs administered in humans resulted in an immune response that produced human anti-mouse antibodies (HAMA). The HAMA response also decreased the retention half-life of mAbs within the body, reducing their effectiveness as a therapeutic. This led to extensive efforts aimed at reducing the immunogenicity of mouse mAbs by, for example, replacing the mouse constant regions of a mAb with human constant regions (Yamada, 2011). At the same time, the development of phage display bypassed the hybridoma technology and allowed high-throughput screening of high affinity binders along with the creation of large human antibody repertoire screening libraries (Mondon et al., 2008). Phage display works by replacing the sequence of the phage coat protein with the sequence of a mAbs V region, which results in a fusion protein that preserves the infectivity of the phage while allowing the immunoglobulin to

be displayed on the surface, bind its target antigen and be affinity purified as a result (McCafferty et al., 1990; Smith, 1985).

The next major advancement in humanising mAbs came from genome engineering approaches and the creation of humanised transgenic mice. The use of human immunoglobulin sequences in the production of therapeutic antibodies circumvents the potential immunogenic responses in patients to antibodies raised using other species intrinsic sequences. However, the use of human constant regions in transgenic mice resulted in inefficient developmental progression with substantial reduction of B cell numbers (Brüggemann et al., 1989; Fishwild et al., 1996; Green et al., 1994; Lonberg et al., 1994). This may be a result of inefficient interfacing of the human immunoglobulin transmembrane constant region with the murine Igα and Igβ signal transducers at the pre-BCR and BCR stages of development. Nucleotide comparison of the Igh transmembrane domains showed some differences between human and mouse (Varriale et al., 2010), supporting this idea. As a result, the next iteration of transgenic mice used the endogenous constant region in combination with human V, D and J genes to obtain potent *in vivo* affinity matured chimeric mAbs (Lee et al., 2014; Ma et al., 2013; Macdonald et al., 2014; Murphy et al., 2014; Osborn et al., 2013). The mAbs can have the murine constant region re-engineered post hoc, providing fully humanised mAb therapeutics.

Despite the extensive pursuit of mAb therapeutics, mAbs have several limitations. They are large in size, difficult to engineer and relatively expensive to manufacture. In addition, due to their multimeric structure they can be relatively unstable (Chames et al., 2009; Steels and Gettemans, 2018). The derivation of small fragments from antibodies, that can retain the binding activity of the original molecule, was a significant advance and gave rise to new antibody formats. The first demonstration of such moieties was the scFv fragment composed of only the heavy and light chain variable region from a mAb linked together with a flexible linker (Bird et al., 1988). Subsequently, Greg Winter's group demonstrated that *E. coli* can produce mouse heavy chain variable domains without the light chain while retain their binding capacity, and named these fragments single domain antibodies (dAbs) (Ward et al., 1989). The next innovation came with the discovery of diabodies by the same group, improving on the short half-life of scFv fragments in serum and overcoming the aggregation of dAbs in solution (Holliger et al., 1993). Diabodies lack a long linker that results in two V_H:V_L domains forming a dimer and because of the two unique V_H:V_L components, diabodies can form bispecific interactions. The modular nature of fragment antibodies and their ease of manufacture in prokaryotic systems means that they can be used in a multi-array configuration with other effector proteins fusions such as cytokines, immunotoxins and membrane proteins to create chimeric receptors. More recently, naturally occurring dAbs, called heavy chain only antibodies, were first discovered in the serum of camels and later shown in other animals that are part of the Camelids group (camels, llama, vicugna)

in addition to cartilaginous fish such as nurse sharks (Flajnik and Kasahara, 2010; Hamers-Casterman et al., 1993).

1.7 Heavy chain only antibodies (HCAbs/nanobodies)

Heavy chain only antibodies (HCAbs) differ from conventional antibodies in two respects: they are missing the C_H1 domain and they do not have a light chain (LC) (Hamers-Casterman et al., 1993) (Figure 1-8). The heavy chain C_H1 domain normally forms disulfide bonds with LC and its absence is the primary cause of the LCs inability to bind the camelid variable domains (VHHs). The VHH of an HCAb is the smallest part of an antibody (around 15kDa) that is able to strongly and specifically recognise its target (Figure 1-8) (Steels and Gettemans, 2018). The VHHs of camelids have been shown to have low propensity for aggregation and an ability to refold after high temperature denaturation (Ewert et al., 2002; Pérez et al., 2001), providing a clear advantage over earlier dAbs. The higher solubility of the VHH domains has been attributed to the substitution of certain residues within the FR2 region. Alteration of these residues in a human variable domain (VH) leads to the same higher solubility observed in camelid VHH, confirming their importance (Davies and Riechmann, 1994). These residues are normally involved in hydrophobic interactions between V_H and V_L chains and their conversion into more hydrophilic residues allows VHHs to further resist pairing with the V_L (Genst et al., 2006; Steels and Gettemans, 2018). The lack of a V_L in HCAbs results in a lower combinatorial diversity of camelid VHHs. However, on average camelid VHHs have a longer CDR_H3 domain with an overall higher conformational variation compared to conventional human and mouse antibodies (Wu et al., 1993; Muyldermans et al., 1994). The VHH CDR_H3 domain also features an additional disulphide bond that is thought to stabilise the domain and allow novel loop conformations (Nguyen et al., 2000) (Figure 1-8). The longer CDR_H3 region increases the VHHs antigen interacting surface area and compensates for the lack of V_L, which along with structural variation increases the diversity or the VHH repertoire. The appearance of longer CDR_H3 domains is thought to be the result of selection for properly folded domains and higher specificity antigen binding VHHs (Muyldermans, 2013).

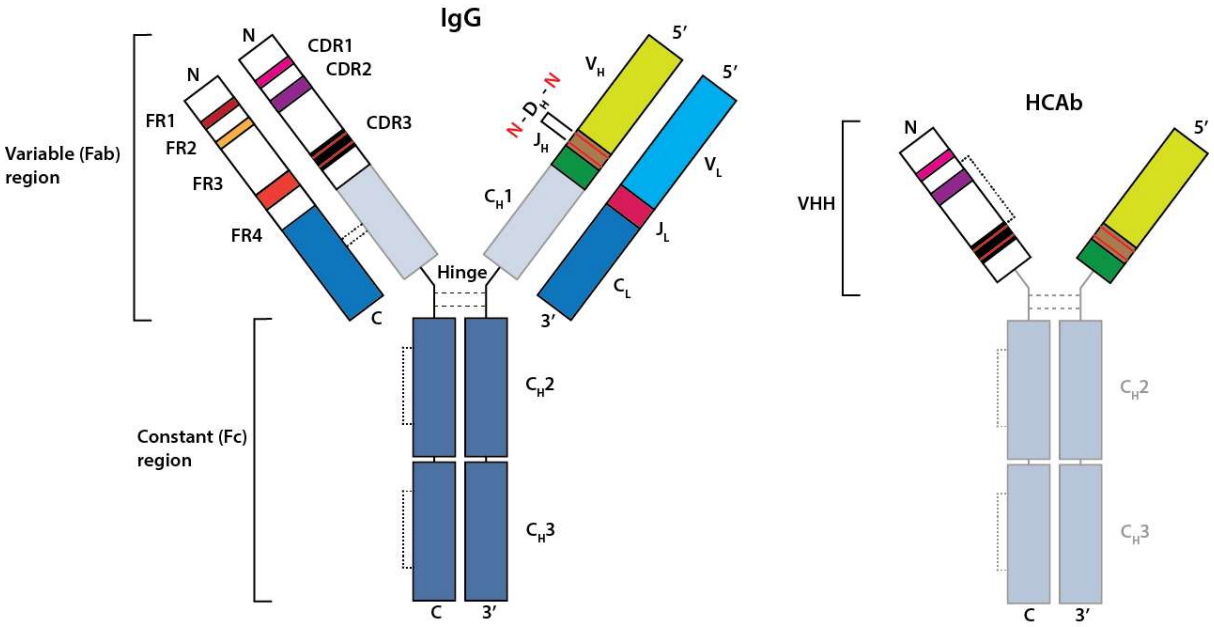


Figure 1-8: The structure of a conventional IgG antibody (left) alongside a heavy chain only antibody (right). One of the key differences is the lack of the C_H1 domain in the HCAb. It is through this domain that the immunoglobulin light chain is covalently attached to the heavy chain. The left portion of each diagram shows protein annotations; the dotted lines represent disulphide bonds. The right portion of the diagram shows DNA sequence annotations. The V_HH fragment (V_H for human HCAb) of the HCAb is the smallest part of the antibody that is still capable of specifically binding an epitope of an antigen. Dashed lines represent disulphide bonds. Adapted from (Georgiou et al., 2014).

1.7.1 Therapeutic applications of HCAbs

HCAbs provide several key advantages that overcome the limitations of mAbs and provide additional functionality. The small size of HCAbs endows them with superior tissue penetrance that is especially critical in cancer therapy, while also allowing them to target conventionally inaccessible surfaces (clefts) such as the active sites of enzyme. This ability allows HCAbs to be used as inhibitors and disruptors of protein interactions. Much like other antibody fragment formats, HCAbs inexpensive manufacture, ease of engineering and higher stability have made them into attractive therapeutics, with multiple clinical trials currently underway. Dimeric or trimeric HCAbs created with flexible linkers can improve avidity or create bispecific binders that improve the HCAbs binding specificity, a desirable feature in therapeutic applications such as tumour targeting (Bannas et al., 2017). In addition, multimeric HCAb formats are able overcome their relatively short half-life *in vivo*, while their ability to retain function within the reducing environment of cells has made them a powerful research tool (Steels and Gettemans, 2018). Multimeric HCAb formats have an increased size (30-45 kDa) compared to the monovalent HCAb (15 kDa) that can reduce their penetration efficiency; however, their size is still much smaller than the 150 kDa of mAbs (Bannas et al., 2017).

Early studies comparing camelid VHH domains to human VH domains have shown camelid VHHs have more desirable biophysical properties, such as reversible folding and the ability to resist aggregation after denaturation (Ewert et al., 2002), suggesting human VH would make less desirable nanobodies.

However, later studies have identified and selected human VH domains that had the same biophysical properties as camelid VHHs (Jespers et al., 2004). Within the enormous human antibody repertoire, it is not surprising that a subset of human VH domains would possesses camelid VHH qualities. Studies trying to pinpoint the consensus of this subset have found some amino acid commonalities (Christ et al., 2007; Dudgeon et al., 2009). However, no large-scale studies with more than a few variable region sequences have been conducted to determine if this subset of human VHs arises from specific VDJ gene usage or through downstream random diversification processes such as TdT nucleotide incorporation and somatic hyper-mutation (SHM). The ability to use human VDJ genes for the production of HCAbs is key in preventing the production of human anti-mouse antibodies (HAMA response) and therefore enabling their use as therapeutics.

2 Methods

2.1 Animals

C57Bl/6JBabr mice (hereafter referred to as wildtype (WT)) were housed and bred at the Biological Support Unit (BSU) at the Babraham Institute, Cambridge. Crescendo mice (humanised HCAb producing transgenic mice created by Crescendo Biologics) were housed and bred at the Charles River laboratories facility (Manston, Kent, UK). WT mice were culled using carbon dioxide asphyxiation in accordance with the Humane Killing of Animals under Schedule 1 to the Animal (Scientific Procedures) Act 1986. Crescendo mice legs and spleens were shipped over night or on the same day in RPMI-1640 media (cat. no. R8758; Sigma) supplemented with 5% foetal calf serum (FCS) (cat. no. F9665; Sigma).

The following text has been redacted due to sensitivity of the material

2.2 Cell lines

Nalm6 (B cell precursor leukaemia cell line) cells were grown in RPMI-1640 (cat. no. R8758; Sigma) supplemented with 10% FCS (cat. no. F9665; Sigma) and Penicillin Streptomycin (cat. no. 15140-122; Invitrogen; 100x) at a 0.5x final concentration. For storage the Nalm6 cells were frozen down in 70% RPMI-1640, 20% FBS and 10% DMSO. All media was filter sterilised before use (cat. no. 431097; Corning). Nalm6 cells were maintained at 1-2 million/ml density, split 1:2-1:3 every 3 days and seeded at 1 million/ml density (minimum 0.5 million/ml).

Yeast with 10V YAC construct were grown in synthetic complete drop out medium composed of 1.725 g of SP supplements (cat. no. CYN0405; Formedium), 0.18 g amino acids mix (cat. no. DCS0741; Formedium), 200 ml water and 4 g glucose (cat. no. G/0500/53; Fisher Scientific) which was added after autoclaving at 121 ^{o}C for 15 minutes.

The naïve and primed hPSC were cultured by Amanda J. Collier, under the supervision of Peter Rugg-Gunn. All work was carried out in accordance with the UK Stem Cell Bank Steering Committee and Babraham Institute Health and Safety and Human Tissue Ethics Committees.

2.3 3D DNA Fluorescent *in-situ* hybridization (FISH)

3D DNA FISH was performed as previously described (Bolland et al., 2013). Briefly, 50-20 µl of cells at a concentration of 10 million cells/ml were placed onto poly-L-lysine coated slides (cat. no. P0425; Sigma) and left for 5 minutes to allow attachment. The slides were placed into 4% formaldehyde/PBS (PFA; cat. no. 28794.295; VWR) for 10 minutes and quenched in glycine for at least 10 minutes at 22^{o}C.

After permeablisation using Saponin (cat. no. 4521; Sigma), the slides were stored in 50% glycerol at -20 °C for at least a week before proceeding to three rounds of freeze thawing in liquid nitrogen. After PBS (cat. no. 70011044; Gibco) washes, the slides were placed in 0.1N HCl, rinsed and then subjected to another round of permeablisation using Saponin and Triton X-100 (cat. no. T9284; Sigma). Probes (Table 2-1) precipitated with mouse Cot-I (cat. no. 18440016; Invitrogen) and salmon sperm DNA (cat. no. D7290; Sigma) were added to the cells which subsequently underwent denaturation at 78°C for exactly 2 minutes, followed by a 16 hr incubation at 37 °C to allow hybridisation. After SSC washes, cells were counterstained with DAPI followed by an additional fixation with 3.7% formaldehyde, which results in a cleaner signal and longer storage life. Finally, a coverslip was mounted along with a drop of Vectashield (cat. no. H-1000; Vector Laboratories) or ProLong Diamond antifade mounting media (cat. no. P36970; Invitrogen). Imaging was performed on the Olympus BX61 fluorescent microscope or on the Metafer/MetaCyte slide scanning system with the help of Simon Walker, Daniel Bolland, Jannek Hauser and Fatima Santos. In general, 13 Z-stacks were imaged and maximum intensity merged using ImageJ v1.48 (Schneider et al., 2012).

Table 2-1: FISH probes used to visualise the IgH locus and the Crescendo inserted YAC transgene.

Probe contains	**Species**	**Labelling**	**Origin**
IgH constant region – all constant genes	Mouse	Alexa fluor 488 (cat. no. A32750; Invitrogen)	RP24-258E20 BAC
VD intergenic region probe 1-7	Mouse	Alexa fluor 555 (cat. no. A32756; Invitrogen)	PCR product (Ramadani et al., 2010)
VD intergenic region probe 1-4	Human	Alexa fluor 555	PCR product (see Table 2-2)
Cγ, Cε, Cα, 3'RR	Mouse	Alexa fluor 488	RP23-149L24 BAC
J, Cμ, Cδ, Cγ genes	Mouse	Alexa fluor 488	RP23-109B20 BAC
IGHV1-2 – IGHV3-13	Human	Alexa fluor 647 (cat. no. A32757; Invitrogen)	RP11-659B19 BAC
IGHV3-21 – IGH4-39	Human	Alexa fluor 647	RP11-72N10 BAC
Chr14 paint	Mouse	Green	CytoCell AMP14G-S
Chr12 paint	Mouse	Green	CytoCell AMP12G-S

2.3.1 DNA FISH probe design and generation

DNA FISH probes for the human IGH V-D intergenic region were designed using the web engine boosted fluorescence in-situ hybridisation (webFISH) tool (Nedbal et al., 2012) (Table 2-1). The human genome assembly GRCh37 was used. Probe design was limited to a small region on chromosome 14 with coordinates 106385388-106405563 that corresponds to the V-D intergenic region. The tool maximises coverage of repetitive regions such as the IgH by also considering locally repetitive regions.

These regions are absent from the rest of the genome, hence making them good probe targets despite their repetitiveness.

The 10V YAC yeast were grown until OD 0.69 and collected with a 2975 xg spin for 2 minutes. DNA was extracted using the YeaStar Genomic DNA kit (cat. no. D2002; Zymo Research). The webFISH-generated primers were used to amplify the VD intergenic regions from the extracted DNA. Gel extracted PCR products were inserted into the pGEM-T Easy Vector plasmid and expanded up in DH5α cells. The probe sequence was verified using Sanger sequencing.

For probes made from BACs, the BAC DNA was extracted using the NucleoBond BAC 100 kit (cat. no. 740579; Macherey-Nagel), following manufacturers protocol.

2.3.2 Nick translation probe labelling

Nick translation labelling of probes was performed as previously described (Bolland et al., 2013) with some modifications. Briefly, mixed on ice 5 µl 10x NTB (0.5 M Tris-HCl pH 7.5, 50 mM $MgCl_2$, 0.5 mg/ml nuclease-free BSA fraction V), 0.1 M DTT, 4 µl d(GAC)TP mix 0.5 mM, 1µl dTTP 0.5 mM, 6 µl aminoallyl-dUTP 0.5 mM (cat. no. R1101; Thermo Scientific), 1 µl DNA Polymerase I 10 U/µl (cat. no. M0209; NEB), 1µl DNase I dilution (1:25) (cat. no. 04716728001; Roche), H_2O to a final volume of 50 µl. The mix was incubated for exactly 2 hours at 16 °C and then for 5 minutes at 75 °C to inactivate DNase I. DNA digestion was visualised using a 2% agarose gel. Each probe was nick translated in a separate reaction (1 µg each). Optimally digested DNA was pooled (4 µg), ethanol precipitated and the pellet was re-suspended in 4 times 1.25 µl H_2O followed by quantification using the NanoDrop spectrophotometer (cat. no. ND-1000; Thermo Fisher Scientific). Single-use dried pellet of amine reactive dye was re-suspended in 2 µl anhydrous DMSO (cat. no. 154938; Sigma). For a 10 µl reaction, 2 µg of DNA in a 5 µl volume was heated to 95 °C for 5 minutes and placed on ice. 3 µl of NaB buffer (0.2 M Sodium bicarbonate pH 8.3) was added to snap cooled amine-modified DNA followed by 2 µl of re-suspended dye. After 1 hour incubation the QIAquick PCR purification kit (cat. no. 28104; Qiagen) was used to purify the probes. The probe florescent intensity was analysed on the NanoDrop 1000.

Table 2-2: Primers obtained from webFISH and used for VD intergenic region probe generation.

Name	Forward primer	Reverse primer
U1	TGCAACGTGCCCTTGTAACACC	AGAACCCACACCATATTCTTTGACTTGTGC
U2	CCAACAGCTCACCTGCAGCC	TGGCACCTCTTAAATCCCTGATCTTGC
U3	GCTTCCCAACTCATTCTGTGAGTCCAGC	GAAGAGGGACCAGGTTGGGAGGC
U4	TCAGCCAATTTAAGGAGGTTTTCCAGTTGC	TGCTCAGGACCCCAGAAGGC

2.4 General molecular biology methods

All solution and general molecular biology methods were performed as described before (Sambrook and Russell, 2000).

2.4.1 PCR agarose gel extraction

The 1-2% agarose gel prepared with TBE and stained with 40 µg/l ethidium bromide were run alongside a lane containing MassRuler DNA ladder (cat. no. SM0403; Thermo Scientific) and visualised using the Molecular Imager Gel Doc XR System from Bio-Rad (Model: Universal Hood II). PCR products were excised using a UV fluorescent table and a razor. Slices were placed into a pre-weighted 1.5 ml Eppendorf tube. The PCR products were purified from agarose using QIAquick Gel Extraction kit (cat. no. 28704; Qiagen) according to manufacturer's instructions and quantified on the NanoDrop.

2.4.2 Cloning

Cloning was performed using the pGEM-T Easy Vector System I from Promega. 25 ng of pGEM-T Easy vector was incubated with T4 DNA ligase and a ratio of less than 1 to 8 vector to PCR product for 30 minutes at room temperature and then at 4 ^{o}C overnight or at 16 ^{o}C overnight. 30 µl of subcloning efficiency DH5α chemically competent *E.coli* (cat. no. 18265017; Invitrogen) were transformed with the ligated vector as per manufacturer's instructions. The transformed *E.coli* were incubated at 37 ^{o}C for 1.5 hours in 100 µl of S.O.C media (cat. no. 15544034; Invitrogen) on a Thermomixer (cat. no. 12799008; Eppendorf). Petri dishes containing Luria-Bertani (LB) Agar with 100 µg/ml of Ampicillin were prepared and used to plate out transformants.

2.4.2.1 Screening transformants

Single colonies were picked, after overnight incubation at 37 ^{o}C, and used for starter cultures composed of 3 ml LB broth with 100 µg/ml Ampicillin. Part of the pick colony was dipped into a PCR mix to screen for successfully transformed clones.

2.4.3 Sanger DNA sequencing

Sequencing of plasmids was performed by Beckman Coulter Genomics using the T4 universal primers. The results were visualised and analysed using ApE (A plasmid Editor - biologylabs.utah.edu/jorgensen/wayned/ape/).

2.4.4 Primer design

All primers were designed using Primer-BLAST (Ye et al., 2012). For large arrays of primers, the primers were designed using a custom python script wrapper for Primer3 (Rozen and Skaletsky, 2000).

2.5 Transgene localisation using a restriction enzyme-based PCR method

The transgene amplification with adjacent genomic sequence was performed as previously described (Bryda and Bauer, 2010), with modifications. Genomic DNA was extracted from CD19 positive bone marrow B cells using the DNeasy Blood & Tissue kit (cat. no. 69506; Qiagen) following manufacturer's instructions. The DNA was digested for 2 hours at 37 ^{o}C using five restriction endonucleases: BglII, EcoR1, PstI, SacI, XbaI. All PCR reactions were performed with Qiagen HotStarTaq DNA polymerase (cat.no. 203203; Qiagen) using the Q-solution (Betaine) which helps in the amplification of complex

templates (rich in GC) by altering the DNA melting temperatures. The first single round of PCR was performed using the Biometra T3 thermocycler with the 3R primer (Table 2-3) at 95 °C for 15 minutes, 60°C for 1 minute, 72 °C for 10 minutes, and hold at 9 °C. After preparation of the Y linker (annealing of Y linker A and E) (Table 2-3), the PCR product was incubated for 16 hours at 16 °C with T4 DNA ligase (cat. no. M0202M; NEB). The second round and third round of PCR was performed using 2R and 1R primers (Table 2-3) and the resulting PCR products were run out on a 1 % agarose gel alongside a lane containing MassRuler DNA ladder (cat. no. SM0403; Thermo Scientific). The distinct band above 848bp obtained from XbaI digested DNA was excised and extracted. The extracted DNA was cloned using the pGEM-T Easy Vector System I from Promega and sequenced.

Table 2-3: Sequences of primers and linkers used for transgene localisation.
Primers 1R-3R were used to extend outwards into the unknown sequence adjacent to the transgene. Primers D and G were used in conjunction with the Y linker to extend inwards from the unknown sequence into the transgene.

Name	Forward primer	Reverse primer
3R		TGAGCGAGGAAGCGGAAGAGCGCCTG
2R		GCACTCTCAGTACAATCTGCTC
1R		ACACCCGCCAACACCCGCTGAC
Primer D	GCAAACGATAAATGCGAGGACGGT	
Primer G	ATGCGAGGACGGTACACGGCGACC	
Y Linker A	GTGCAGCCTTGGGTCGCCGTGT (3'-Spacer-C3-CPG)	
Y Linker E	GCAAACGATAAATGCGAGGACGGTACACGGC GACCCAAGGCTGCACT	

2.6 Targeted locus amplification (TLA)

TLA was performed as previously described (Vree et al., 2014). 10 million non-recombining cells from a bone marrow depletion using Ly6c, Ter119, Mac1, Gr1 biotinylated antibodies (see Table 2-8) were collected for TLA. After a spin (250 xg for 10 minutes at 22 °C), the cells were resuspended in 1ml PBS/10% FCS and fixed for 10 minutes at 22 °C with the addition of 37 % formaldehyde (cat. no. 1039991000; Merck) to a final concentration of 2 %. 1 M glycine was added and tubes were placed on ice followed by a PBS wash (500 xg for 2 minutes at 22 °C) and the addition of 500µl lysis buffer (50 mM Tris pH 7.5; 150 mM NaCl; 5 mM EDTA; 0.5 % NP-40; 1 % Triton X-100). Lysis was performed 5 minutes at 22 °C followed by 5 minutes at 65 °C and 1 minute on ice. After lysis, the supernatant was removed (1000 xg for 2 minutes at 22 °C) and the pellet was washed in 400 µl 1x NEBuffer 4 (cat. no. B7004; NEB) (1000 xg for 2 minutes at 22 °C), and resuspended in a final volume of 300µl 1x NEBuffer 4. 12 µl of 5 % SDS was added and the sample was incubated at 37 °C for 30 minutes while shaking (900 RPM), followed by the addition of 30 µl 20 % Trition X-100 and further incubated at 37 °C for 30 minutes while shaking (900 RPM). 400 U of NlaIII (cat. no. R0125; NEB) were added and incubated overnight at 37 °C while shaking (900 RPM). Digestion efficiency was determined by running a digested and undigested control on a 0.6 % agarose gel. The restriction enzyme was inactivated by incubation

at 65 °C for 20 minutes after which 40 U of T4 DNA ligase (cat. no. 15224025; Invitrogen) were added and incubated with the nuclei for at least 2 hours at 22 °C. De-crosslinking was performed with 5µl Proteinase K (10 mg/ml; cat. no. 03115879001; Roche) overnight at 65 °C, followed by 5µl RNase A (10 mg/ml; cat. no. 19101; Qiagen) digestion at 37 °C for 10 minutes. A phenol/chloroform (cat. no. P3803; Sigma) purification was performed and the aqueous phase was ethanol precipitated with 3 M Sodium Acetate pH 5.2 (1/10 volume) and 100 % Ethanol (1.8x volumes) at -80°C until sample was completely frozen. The DNA was spun down (20,000 xg for 20 minutes at 4 °C), supernatant discarded and pellet washed with 70 % ethanol followed by another spin (20,000 xg for 4 minutes at 22 °C). The air-dried pellet was dissolved in 150 µl 10 mM Tris pH 7.5 at 37 °C. A second restriction enzyme digestion was carried out overnight at 37 °C with NspI (cat. no. R0602; NEB), followed by enzyme inactivation at 65 °C for 25 minutes. 100U of T4 DNA ligase (cat. no. M0202; NEB) was added to the digested DNA, to a final volume of 14 ml, and incubated overnight at 16°C. Next day, the DNA was ethanol precipitated with the addition of 3 M Sodium Acetate (1/10 volume), 14µl Glycogen (20mh/ml; R0551; Thermo Scientific), 100 % ethanol (2.25 volumes) and stored at -80°C until completely frozen. After a spin (3200 xg for 15 minutes at 20 °C), the supernatant was discarded and the pellet was washed with 70% ethanol, followed by another spin (3200xg for 15 minutes at 20°C). The pellet was air-dried and dissolved in 150 µl 10 mM Tris pH 7.5 at 37 °C. The TLA PCR (34 cycles, 2 minutes each) was performed in 25 µl reactions with 100ng of DNA using anchor primers that are facing away from each other (Table 2-4) and the Phire polymerase (cat. no. F122S; Thermo Scientific). The concentration and fragment length profiles of the TLA library was determined with the 2100 Bioanalyzer (cat. no. G2939BA; Agilent). Library quantity was determined using the KAPA Illumina Library Quantification Kit (cat. no. KK4824; KAPA Biosystems). The library was 150 bp pair-end sequenced on the Illumina MiSeq system.

Table 2-4: TLA anchor primers used to amplify circularised DNA from within the YAC transgene arm on the J gene proximal side.

The following table has been redacted due to sensitivity of the material

2.6.1 Transgene analysis primers

Long amplicon PCR was performed with either the Phire polymerase (cat. no. F122S; Thermo Scientific) or the Q5 high-fidelity polymerase (cat. no. M0492S; NEB).

Table 2-5: Primes used to check the integrity of the transgene YAC arms.

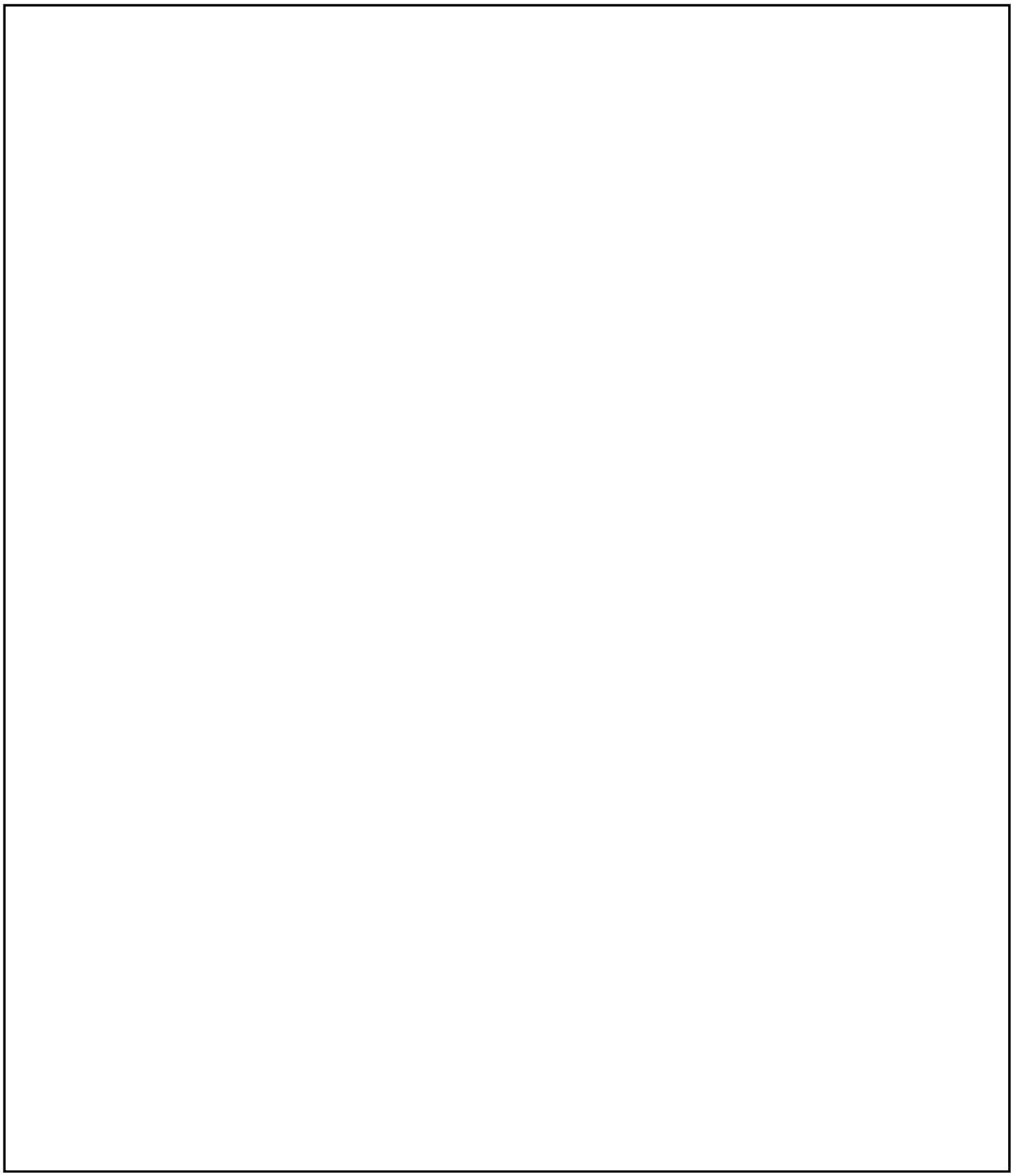
The following table has been redacted due to sensitivity of the material

Table 2-6

The following table has been redacted due to sensitivity of the material

Table 2-7

The following table has been redacted due to sensitivity of the material



2.6.2 TLA analysis

For genome tracks and read count plots, the TLA reads were aligned to GRCm38 to determine the location of the transgene and to GRCh37 to uncover the composition of the transgene. Bowtie 2 (Langmead and Salzberg, 2012) with `--very-sensitive --score-min L,0,-1` settings was used for alignment. Reads that did not align the first time were *in-silico* digested with NlaIII and aligned again. BAM files were merged (Li et al., 2009) and analysed using Seqmonk v1.42.0 (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). GRCm38 blacklisted regions (The ENCODE Project Consortium, 2012) were excluded from the analysis. Ensembl GRCh37 annotation

version 87 was used for IGH gene coordinates. For the GRCm38 alignment, 12kb-sized probes were used to visualise the data.

The following text has been redacted due to sensitivity of the material

Aligned sequences were visualised with the Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

The following text has been redacted due to sensitivity of the material

2.7 Sample preparation for FISH and flow cytometry

2.7.1 Cell isolation from bone marrow

Bone marrow was extracted from femurs and tibias by removing the ends of bones and flushing with RPMI-1640 media (cat. no. R8758; Sigma) supplemented with 5 % FCS (cat. no. F9665; Sigma) and 25 mM HEPES (cat. no. 15630056; Gibco) (bone marrow media) using a 25-gauge needle and 2 ml syringe. A 10 ml stripette was used to dissociate extracted bone marrow, after which the content was transferred into a fresh 50 ml falcon, leaving bone fragments behind.

2.7.2 Cell isolation from spleen and thymus

C57BI/6JBabr spleens/thymus were extracted and stored in bone marrow media. The spleens/thymus were forced through a 70 µm cell strainer (cat. no. 352350; BD) using the round bottom of a 1.5 ml Eppendorf tube and with regular additions of bone marrow media.

2.7.3 Cell preparation

Cells from bone marrow, spleen and thymus were prepared in the same manner. Cells were spun at 336 x*g* 4 °C for 5 minutes and bone marrow medium was decanted. The cells were re-suspended in ice cold Dulbecco's PBS (D-PBS; cat. no. D8537; Sigma) and spun at 448 x*g* 4 °C for 8 minute to burst the erythrocytes (red blood cells) by osmotic lysis. Cells were re-suspended in 10ml of MACS buffer (2 mM EDTA, 0.5 % FCS in PBS) and filtered through a 70 µm cell strainer (cat. no. 352350; BD) with the use of the round bottom of a 1.5 ml Eppendorf tube to push through any residual tissue. A sample of cells was taken, diluted 1:10 and counted on a haemocytometer.

2.7.4 Depletion using magnetic cell sorting (MACS)

Cells prepared and counted with the haemocytometer were re-suspended in MACS buffer (2 mM EDTA, 0.5 % FCS in PBS) at $25x10^6$ cells/ml with biotin conjugated antibodies (see Table 2-8). After a 30 minute incubation on ice, the cells were diluted with D-PBS, spun 336 x*g* at 4 °C for 5 minutes and re-suspended at a concentration of 10^7 cells per 95 µl of MACS buffer. 5 µl of streptavidin MACS beads (cat. no. 130-048-101; Miltenyi) was added and cells were incubated for 15 minutes at 4 °C, with periodic stirring. The cells were filtered through a 40 µm cell strainer (cat. no. 352340; BD) to prevent the LS columns (Miltenyi Biotech 103-042-401) from clogging. A single LS column was prepared at 4 °C

in the cold room by placing it onto the provided magnetic stand and washing it with 3 ml of MACS buffer. The streptavidin incubated cells were washed, spun (336 x*g* for 5 minutes at 4°C) and the pellet was disrupted with 4-10 ml MACS buffer, depending on cell count. The cell suspension was placed onto the LS column (1-3 mice 1 column; 4-7 mice 2 columns; 8-10 mice 3 columns; 11-13 mice 4 columns) and the flow through was collected as the depleted fraction. Two 1 ml washes with MACS buffer were performed to collect any residual cells. The column was taken off the magnet and the undesired enriched cells were washed off with 5 ml of MACS buffer using the provided plunger (these were used for certain assays such as TLA, but were often discarded).

Table 2-8: List of antibodies used for magnetic cell depletion of cells other than the desired B cells. This was primarily performed to reduce FACS times.

Antigen	Clone	Dilution	Source	Expressed on
Ly6c	ER-MP20	1:400	MCA2389B Bio-Rad	Granulocytes; macrophage/dendritic cell precursors; subpopulations of B- and T-lymphocytes; endothelial cells
Ter119	TER-119	1:400	13-5921-85 Ebioscience	Erythroid cells
CD3e	eBio500A2 (500A2)	1:800	13-0033-81 Ebioscience	Thymocytes; mature T cells; NKT cells
Mac1 (CD11b)	M1/70	1:1600	13-0112-82 Ebioscience	Macrophages; NK cells; granulocytes; activated lymphocytes; B-1 cells
Gr1 (Ly-6G)	RB6-8C5	1:1600	13-5931-81 Ebioscience	Monocytes; granulocytes; neutrophils

2.8 Flow cytometry

2.8.1 Staining

All cells were stained at $20x10^6$ cells/ml concentration in MACS buffer.

2.8.1.1 Titration

All antibodies were titrated prior to use. The optimal concentration was calculated by dividing the median fluorescent intensity (MFI) of the positive population by the MFI of the negative population (set using an unstained control) giving a signal-to-noise ratio. The concentrations with the highest signal to noise ratio were used (Table 2-9).

Table 2-9: Antibodies used in flow cytometry cell surface staining of B cell subpopulations.

Antigen	Conjugate	Clone	Dilution	Source
B220	BV421	RA3-6B2	1:200	Biolegend UK
B220	PerCP-Cy5.5	RA3-6B2	1:400	BD Pharmingen
CD19	PerCP-Cy5.5	1D3	1:800	BD Pharmingen
CD43	FITC	S7	1:200	BD Pharmingen
CD43	PE-Cy7	S7	1:1000	BD Pharmingen
BP-1	PE	BP-1	1:200	BD Pharmingen
CD25	APC	PC61.5	1:1000	eBioscience
CD25	Biotin	7D4	1:400	BD Pharmingen
IgM	PE	eB121-15F9	1:400	eBioscience

IgG1	PE	A85-1	1:1000	BD Pharmingen
CD24 (HSA)	Biotin	M1/69	1:400	BD Pharmingen
CD24 (HSA)	FITC	30F1	1:800	eBioscience
CD24 (HSA)	PE	30F1	1:1600	eBioscience
AA4.1	APC	AA4.1	1:2000	eBioscience
CD5	PE-Cy7	53-7.3	1:1000	eBioscience
CD21/35	Biotin	7G6	1:2000	BD Pharmingen
CD23	BV711	B3B4	1:400	BD Pharmingen

Table 2-10: Streptavidin conjugated flurochromes used as the second layer to label biotin conjugated antibodies.

Second layer	Dilution	Source
SA-BV605	1:500	BD Pharmingen

Table 2-11: Live dead stains used

Live/dead stain	Dilution	Source
FVD eflour 506	1:1600	eBioscience
FVD eflour 780	1:2000	eBioscience
DAPI	Final concentration - 300 nM	Roche

2.8.2 Staining panels

2.8.2.1 Bone marrow

An antibody panel with different fluorochromes was used to analyse the bone marrow B cell developmental stages (Table 2-12). The panel was an adaptation of a previously reported gating strategy (Hardy et al., 1991; Rumfelt et al., 2006)(Figure 2-1). This gating strategy and a reduced panel of five colours, with a live dead stain, were used for sorting population for downstream analysis (Table 2-13 and Figure 2-2).

Table 2-12: Flow cytometry bone marrow staining panel

Antigen	Fluorophore
B220	BV421
CD43	PE-Cy7/FITC
CD24	FITC/Biotin SA-BV605
CD19	PerCP-Cy5.5
AA4.1	APC
IgG	PE
Fixable viability dye	eflour 506/DAPI
CD16/32 Fc block	-

Table 2-13: Fluorescence-activated cell sorting bone marrow staining panel

Antigen	Fluorophore
B220	BV421
CD19	PerCP-Cy5.5
CD43	PE-Cy7
CD24	PE
AA4.1	APC
Fixable viability dye	eflour 506/eflour 780

The following figure has been redacted due to sensitivity of the material

(Legend on next page)

Figure 2-1: A representative gating scheme for bone marrow B cells.
The different developmental stages are assigned for the bone marrow staining panel 2 (CTG2 depicted) as defined by the Hardy fractions (Hardy et al., 1991). The dot plot with fraction F, BC and C'DE cells shows the reasoning for the positioning of gates separating fraction D (fraction D and fraction C' are joined as one population and only labelled fraction D here) and E, as well as the final fraction F (B220 high CD43- fraction F cells used in plot). The separation of $CD19^+$ from $CD19^-$ cells was based on the highest density of double negative cells.

The following figure has been redacted due to sensitivity of the material

Figure 2-2: Gating strategy used for sorting fractions BC/C'DE/intermediate EF/F.
Sorting gates are highlighted in blue (CTG2 depicted).

2.8.2.2 Spleen

To examine potential alterations in B cell splenic populations a 7 colour panel with a live/dead stain (Table 2-14) was devised based on previously reported cell surface markers (Allman and Pillai, 2008) with the help of Manuel Diaz-Munoz. The panel allowed enumeration of T1-3, marginal zone (MZ) and follicular (FO) B cell subsets, along with an estimate of B-1 B cell counts (Figure 2-3). The AA4.1 positive staining of T1-3 cells along with the AA4.1 negative staining of MZ and FO cells resulted in a continuous gradient of cells. In order to separate T1-3 from MZ and FO cells an AA4.1 fluorescence minus one (FMO) was run alongside the full panel.

Table 2-14: Flow cytometry spleen staining panel

Antigen	**Fluorophore**
B220	BV421
CD5	PE-Cy7
CD21/35 biotin	SA-BV605
CD23	BV711
AA4.1	APC
IgG	PE
CD19	PerCP-Cy5.5
Fixable viability dye	eflour 506
CD16/32 Fc block	-

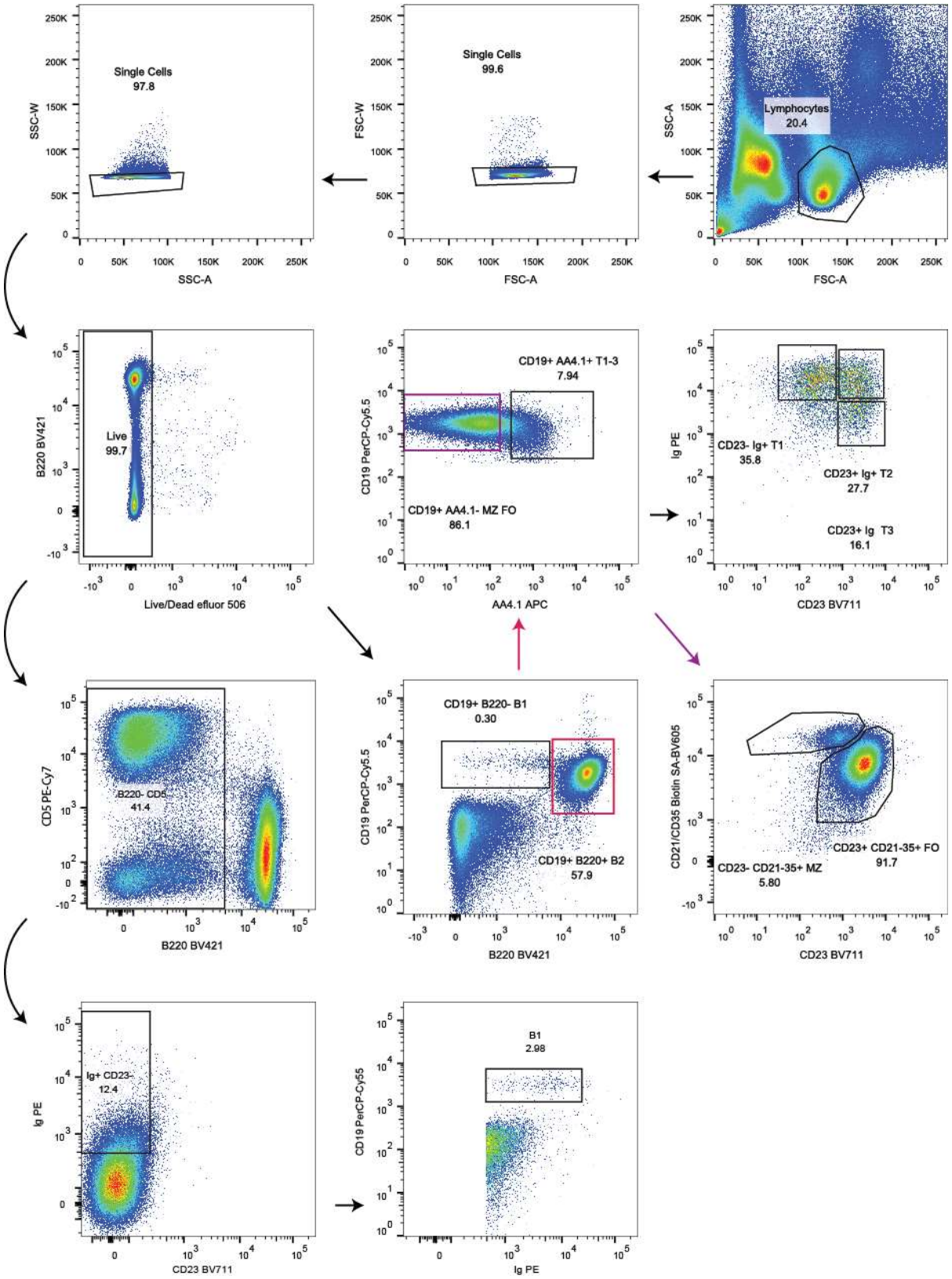


Figure 2-3: Spleen gating strategy for the enumeration of B1, T1-3, marginal zone (MZ) and follicular (FO) B cells. AA4.1 FMO was used to determine were to draw the gate separating T1-3 cells from the MZ FO cell populations (wildtype depicted).

2.8.3 Flow cytometry analysis

Flow cytometry compensation and gating was performed in FlowJ0 10.4. Additional analysis was performed in the R software environment. The FCS files were run through FlowAI v1.2.4 (Monaco et al., 2016) to filter out events affected by technical variation such as abrupt flow rate fluctuations. Subsequently, lymphocytes, singlets and live cells were gated in FlowJo and exported as new FCS files. All data was logicle transformed using the flowCore package (Ellis B et al., 2017), and additionally scaled for FlowSOM analysis. The data was dimensionality reduced for visualisation either using FlowSOM (Gassen et al., 2015) or tSNE (Maaten and Hinton, 2008). FlowSOM uses self-organising maps (SOMs) for dimensionality reduction, allowing it to scale to large numbers of cells. Each cell was assigned to a SOM with a 10x10 grid, which means similar cells are grouped into 100 separate clusters. The clusters were visualised on a minimal spanning tree (MST). For tSNE analysis, the data was downsampled to 10,000 events. Analysis was run multiple times to ensure reproducibility of the results.

2.9 Immunostaining of paraffin embedded spleen sections

Paraffin embedding and spleen sectioning was performed with the instruction of Elena Fineberg and Salah Azzi.

Whole spleens were fixed for 24 hours in 4 % formaldehyde/PBS at 4 °C, followed by tissue dehydration (PBS wash, 30 % ethanol for 30 minutes, 50 % ethanol for 30 minutes, 70 % ethanol for 30 minutes, 100 % ethanol for 30 minutes, 100% ethanol overnight 4 °C; ethanol was diluted with PBS). The dehydrated spleens were embedded in paraffin and left to set overnight. Before sectioning, the blocks were incubated overnight on ice at 4 °C. 0.5 µm sections were cut using a Leica paraffin microtome and mounted onto Superfrost Plus slides (cat. no. 631-0108; VWR). The slides were placed into Histo-Clear II (cat. no. 101412-882; VWR) bath twice for 10 minutes. This removes all the wax from the sections, allowing the spleens sections to be rehydrated (100 % ethanol for 10 minutes, 100 % ethanol for 5 minutes, 70 % ethanol for 5 minutes, 50 % ethanol for 5 minutes, 30 % ethanol for 5 minutes, PBS for 5 minutes or overnight at 4°C; ethanol was diluted with PBS). Epitope retrieval was performed using Sodium citrate buffer (10 mM Sodium citrate, 0.05 % Tween 20, pH 6.0 adjusted with 1 N HCl) boiled for 20-30 minutes in a microwave (800 W; 1 minutes preheat followed by 10 seconds on 50 seconds off for 20-30 minutes).

The epitope retrieved spleen sections were washed in PBS for 5 minutes, followed by 0.5 % Triton X-100 in PBS 10 minutes incubation at 22 °C. Another 5-minute PBS was performed before 100 µl of blocking solution (1 % BSA in PBS) was applied and incubated for 30 minutes in a humidified chamber at 22 °C. The secondary rabbit anti-mouse IgG was diluted 1:100 in blocking solution and 100 µl was applied to each slide. The slides were incubated in a dark humidified chamber for 30-45 minutes,

followed by two 5 minutes washes in 0.1 % Tween 20/PBS and a 5 minutes PBS wash. Sections were counterstained with DAPI/2xSSC for 2 minutes and washed in PBS for 10 minutes. A 3.7 % formaldehyde fixation for 5 minutes was carried with a subsequent 155 mM glycine quench for at least 10 minutes. After a PBS 5 minute rinse, the slides were dried around the spleen section and a coverslip with Vectashield (cat. no. H-1000; Vector Laboratories) was mounted. Imaging was performed on the Olympus BX61 fluorescent microscope.

2.10 VDJ-seq

VDJ-seq of FACS sorted fractions was performed as previously described (Chovanec et al., 2018). Between 1.1-5 µg of DNA was purified using the DNeasy Blood & Tissue Kit (cat. no. 69504, Qiagen). DNA was sonicated to a 500bp average fragment size using the Covaris E220 ultrasonicator (peak incident power (W): 105; duty factor: 5 %; cycles per burst: 200; treatment time (s): 80; temperature (^{o}C): 7; and water level: 6). After end repair and A-tailing the adaptors with a 6 N/12 N UMI were ligated to the DNA fragments using T4 DNA ligase (cat. no. M0202M; NEB). After purification with 1x AMPure XP beads, primer extension was performed using Vent (exo-) (cat. no. M0257; NEB), human biotinylated J gene primers and a single extension cycle. The primer extension products were captured using the MyOne streptavidin T1 Dynabeads (cat. no. 65601; Invitrogen) (5µl of beads for every 1µg DNA) overnight at 22^{o}C while rotating. The beads with the bound primer extension product were washed twice with B&W buffer (10 mM Tris-HCl pH 7.5; 1 mM EDTA; 2 M NaCl; 0.05% Tween 20), once with EB buffer (10 mM Tris pH 8) and resuspended in 10.5 µl for every 1 µg DNA. The DNA (not exceeding 1 µg per 25 µl reaction) was amplified off the beads using the short Illumina adaptors (see Table 2-17) with the Q5 high-fidelity polymerase (cat. no. M0492S; NEB). All reactions were pooled, and the supernatant was collected on a magnetic rack, followed by a 1x AMPure XP bead purification. A second round of PCR amplification was performed with the P5-7 flow cell primers (see Table 2-18) to create the final library, followed by a double 1x AMPure XP bead purification to remove primer dimers created during PCR. The concentration and fragment length profiles of the VDJ-seq libraries were determined with the 2100 Bioanalyzer (cat. no. G2939BA; Agilent). Library quantity was determined using the KAPA Illumina Library Quantification Kit (cat. no. KK4824; KAPA Biosystems). The libraries were 300 bp pair-end sequenced on the Illumina Miseq system.

Table 2-15: Adapter sequences

Name	Sequence (5'-3')
Adapter 1	
P7 adapter F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNGACTCG*T
P7 short adapter R1-block	[phos]CGAGTCNNNNNNAGATCGGAAGAG*C[spcC3]
Adapter 2	
P7 adapter F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNCTGCTCC*T
P7 short adapter R2-block	[phos]GGAGCAGNNNNNNAGATCGGAAGAG*C[spcC3]

* Phosphorothioate bond; short R adapter should be phosphorylated at 5' end and have a carbon-3 spacer incorporated at the 3' end to prevent it being used to prime in later reactions.

Table 2-16: Biotinylated J-specific oligos.

Primer	Sequence (5'-3')
Mouse JH1 Rev Bio	[biotin]AGCCAGCTTACCTGAGGAGAC
Mouse JH2 Rev Bio	[biotin]GAGAGGTTGTAAGGACTCACCTG
Mouse JH3 Rev Bio	[biotin]AGTTAGGACTCACCTGCAGAGAC
Mouse JH4 Rev Bio	[biotin]AGGCCATTCTTACCTGAGGAG
Human JH1 Rev Bio	[biotin]CCAGACAGCAGACTCACCTG
Human JH2 Rev Bio	[biotin]TGCAGTGGGACTCACCTG
Human JH3 Rev Bio	[biotin]AGAAGGAAAGCCATCTTACCTG
Human JH4 Rev Bio	[biotin]CAGGAGAGAGGTTGTGAGGACT
Human JH5 Rev Bio	[biotin]AGGGGGTGGTGAGGACTC
Human JH6 Rev Bio	[biotin]CCATTCTTACCTGAGGAGACG

Table 2-17: Sets of J-specific reverse primers.
The primers map 10 bp for mouse and 15 bp for human downstream of the J junction to allow for nucleotide additions, modifications and deletions at the junction occurring during V(D)J recombination.

Primer	Sequence (5'-3')
P7 short	ACACTCTTTCCCTACACGACGCTC*T
Mouse JH1 P5 short	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CCCTGTGCCCCAGACATCGA*A**
Mouse JH2 P5 short	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**AGTGGTGCCTTGGCCCCAGTA*G**
Mouse JH3 P5 short	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**ACCAGAGTCCCTTGGCCCCAGTA*A**
Mouse JH4 P5 short	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**TGAGGTTCCTTGACCCCAGTAGTCCAT*A**
Human JH1 P5 short	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GGTGCCCTGGCCCCAGT*G**
Human JH2 P5 short	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GGTGCCACGGCCCCAGAG*A**

Human JH3 P5 short	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**ACCATTGTCCCTTGGCCCCA*G**
Human JH4 P5 short	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GACCAGGGTYCCYTGGCCC*C**
Human JH5 P5 short	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CAGGGTTCCYTGGCCCCAG*G**
Human JH6 P5 short 1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CCTTTGCCCCAGACGTCCATGTAG*T**
Human JH6 P5 short 2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**TKSCCCCAGACGTCCATACCG*T**

The primers have part of the Illumina P5 sequence at the 5' end (P5 short), along with the J-specific sequence at the 3' end (in bold). A single forward primer (P7 short) annealing to the P7 sequence in the ligated adapter is used at the other end. Asterisks (*) represent phosphorothioate bonds. Code for degenerate primers: B=G+T+C; D=G+A+T; H=A+T+C; S=G+C; V=A+C+G; W=A+T; K=G+T; M=A+C; N=A+C+G+T; Y=C+T; and R=A+G.

Table 2-18: Flow cell index primers.
Index shown in bold.

Primer	Sequence (5'-3')
P7 flow cell	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC
P5-I5 flow cell index 1	CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 2	CAAGCAGAAGACGGCATACGAGAT**ACATCG**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 3	CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 4	CAAGCAGAAGACGGCATACGAGAT**TGGTCA**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 5	CAAGCAGAAGACGGCATACGAGAT**CACTGT**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 6	CAAGCAGAAGACGGCATACGAGAT**ATTGGC**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 7	CAAGCAGAAGACGGCATACGAGAT**GATCTG**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 8	CAAGCAGAAGACGGCATACGAGAT**TCAAGT**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 9	CAAGCAGAAGACGGCATACGAGAT**CTGATC**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 10	CAAGCAGAAGACGGCATACGAGAT**AAGCTA**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 11	CAAGCAGAAGACGGCATACGAGAT**GTAGCC**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 12	CAAGCAGAAGACGGCATACGAGAT**TACAAG**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 13	CAAGCAGAAGACGGCATACGAGAT**TTGACT**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 14	CAAGCAGAAGACGGCATACGAGAT**GGAACT**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 15	CAAGCAGAAGACGGCATACGAGAT**TGACAT**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 16	CAAGCAGAAGACGGCATACGAGAT**GGACGG**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 17	CAAGCAGAAGACGGCATACGAGAT**CTCTAC**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 18	CAAGCAGAAGACGGCATACGAGAT**GCGGAC**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 19	CAAGCAGAAGACGGCATACGAGAT**TTTCAC**GTGACTGGAGTTCAGACGTGT
P5-I5 flow cell index 20	CAAGCAGAAGACGGCATACGAGAT**GGCCAC**GTGACTGGAGTTCAGACGTGT

2.10.1 VDJ-seq analysis

VDJ-seq was analysed using the BabrahamLinkON pipeline (Chovanec et al., 2018) described in more detail in section 4. For CTG2 mice, the 6bp UMI adaptor was used along with 8bp from the V region to deduplicate all sequences.

The following text has been redacted due to sensitivity of the material

Genes with identical reference sequences were merged (IGHD5-18*01, IGHD5-5^01; IGHD4-11*01, IGHD4-4*01; IGHV1-69*01, IGHV1-69D*01; IGHV3-23*01, IGHV3-23D*01; IGHV3-30*04, IGHV3-30-3*03; IGHV3-30*02, IGHV3-30-5*02; IGHV3-30*18, IGHV3-30-5*01; IGHV3-29*01, IGHV3-30-42*01). Only the first called gene was taken from IgBlast calls that contained multiple V, D, or J gene calls. Sequences with a V gene score of 100 or higher and a J score above 35 were used for analysis. Expected frequency was calculated using the first [redacted] (excluding IGHV7-4-1, IGHV4-30-2 and IGHV3-30-3, which were not in the GCRh37 reference annotation) that should be within the transgene. Multiple testing correction for the binomial test was performed using the Benjamini & Hochberg method (FDR).

VDJ-seq data generated by Louise Matheson (CTG1 and Human VDJ-seq) were processed as previously described (Bolland et al., 2016; Matheson et al., 2017) by Felix Krueger. Quantification was performed in Seqmonk v1.42.0 (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/).

Throughout my thesis, I will be using the IMGT definitions of productive and unproductive to describe rearranged genes and reserve functional, pseudogene and ORF (Open Reading Frame) to describe the germline sequence (Lefranc, 1998). This is an important distinction as functional genes can have both productive and unproductive rearrangements based on the N and P nucleotide addition and nibbling, resulting from non-homologous end joining (NHEJ) at the junctions. Unproductive rearrangements contain a premature stop codon and/or an out-of-frame junction and they do not translate into functional protein. Additionally they may not contain the highly conserved amino acids, such as cysteine (C) at position 23 and tryptophan (W) at position 41 that mark the CDR3 boundaries (Lefranc, 2014).

2.11 CTCF peak proximity to RSS analysis

ENCODE generated ChIP-seq data for the lymphoblastoid cell line GM12878 was used to determine transcription factor proximity to the V gene recombination signal sequence (RSS) (The ENCODE Project Consortium, 2012). Each of the V genes was assigned to one of three main clans based on IMGT nomenclature: clan I comprises of IGHV1, IGHV5 and IGHV7 subgroups; clan II comprises of IGHV2, IGHV4, IGHV6 and IGHV(II) subgroups; clan III comprises of IGHV3 and IGHV(III) subgroups (Lefranc, 2000). V genes not assigned to any of the three clans were labelled as ‘Other’.

2.12 Primed hPSC H9 NK2

H9 NK2 cells were obtained from Austin Smith at the WT-MRC Cambridge Stem Cell Institute with permission from WiCell (Takashima et al., 2014). These cells contain a doxycycline inducible NANOG and KLF2 construct for resetting to a naïve state of pluripotency. The cells were maintained under feeder free conditions on Vitronectin (cat. no. A14700; Thermo Fisher Scientific) coated plates (5µg/ml in PBS and coated for at least an hour at 22°C or overnight at 4°C) in TeSR-E8 (cat. no. 05990; StemCell Technologies) culture medium which was changed daily. The cells were typically passaged at a 1:10 – 1:15 split ratio every 6 days. To passage the cell, the culture medium was aspired and residual media was gently washed away PBS. Cells were incubated for 5 minutes at 22°C with Gentle Cell Dissociation Reagent (GCDR) (cat. no. 07174; StemCell Technologies), followed by aspiration of GCDR and addition of TeSR-E8. The cells were manually dissociated from the culture vessel using a 5ml stripette, collected and passaged to freshly coated Vitronectin vessels.

2.13 Naïve hPSC H9 NK2

H9 NK2 naïve cells were derived from H9 NK2 primed hPSCs after NANOG and KLF2 transgene induction and cultured in t2iL+PKCi media. Fully reset naïve hPSCs were obtained from Austin Smith at the WT-MRC Cambridge Stem Cell Institute with permission from WiCell (Takashima et al., 2014). The N2B27 basal medium for t2iL+PKCi media was prepared as follows: 1:1 mixture of DMEM/F12 (cat. no. 21331020; Thermo Fisher Scientific) and Neurobasal (cat. no. 21103049; Thermo Fisher Scientific); 0.5x N2 supplement (cat. no. 17502048; Thermo Fisher Scientific); 0.5x B27 supplement (cat. no. 17504044; Thermo Fisher Scientific); 2mM L-Glutamine (cat. no. 25030024; Thermo Fisher Scientific); 50U/ml Penicillin-Streptomycin (cat. no. 15140122; Thermo Fisher Scientific); 0.1mM β-mercaptoethanol (cat. no. 31350010; Thermo Fisher Scientific). The complete t2iL+PKCi media was prepared from the N2B27 basal medium with the addition of the following components (used within 2 days of addition): 1µM PD0325901 (WT-MRC Cambridge Stem Cell Institute); 1µM CHIR99021 (WT-MRC Cambridge Stem Cell Institute); 2µM Gö6983 (cat. no. 2285; Tocris); 20ng/ml human LIF (WT-MRC Cambridge Stem Cell Institute). The cells were maintained under feeder free conditions on Matrigel growth factor reduced basement membrane matrix (cat. no. 354230; Corning) coated plates (diluted 1:100 in DMEM/F12 and coated for at least an hour at 22°C or at 4°C overnight) with the t2iL+PKCi media changed daily. The cells were typically passaged every 3-4 days at a 1:4-1:5 split ratio. To passage the cells, the culture media was aspired followed by a 3-5 minute incubation with Accutase (cat. no. A1110501; Thermo Fisher Scientific) at 37°C. Cells were collected and the Accutase was diluted at least 1:1 with DMEM/F12 before a 3 minute 300xg, followed by resuspension in complete medium and passaging onto freshly coated Matrigel vessels.

2.14 Promoter capture Hi-C and Hi-C

Promoter capture Hi-C and Hi-C was performed by Stefan Schoenfelder as previously described (Mifsud et al., 2015; Schoenfelder et al., 2015). Approximately 80x10^6 cells (naïve and primed hPSC) were used to ensure libraries with high enough complexity would be obtained.

2.14.1 Hi-C

Cells were fixed in 2% formaldehyde (cat. no. AGR1026, Agar Scientific) for 10 min and quenched with ice-cold glycine (cat. no. 410225; Sigma) at a final concentration of 125mM. Cells were washed with 50ml PBS pH 7.4 (cat. no. 0010001; Gibco) using a 400xg 10 min 4°C spin, followed by another spin. The cell pellet was snap frozen in liquid nitrogen and stored at -80°C.

After thawing the cell pellet, the cells were resuspended in 50ml ice-cold lysis buffer (10mM Tris pH 8; 10mM NaCl; 0.2% Igepal CA-630; protease inhibitor cocktail (cat. no. 04693159001; Roche) and incubated for 30 minutes on ice. The nuclei were pelleted at 650xg for 5 minutes 4°C and washed with 1.25x NEBuffer 2 (cat. no. B7002; NEB). After the wash, the nuclei were resuspended in 1.25x NEBuffer 2 (7x10^6 cells per Eppendorf), 0.3% final concentration SDS (cat. no. V6551, 10% stock; Promega) and incubated at 37°C for 1 hour with agitation (950 rpm). Triton X-100 (cat. no. T8787; Sigma) was added to the nuclei to a final concentration of 1.7% and further incubated for 1hr at 37°C with agitation (950 rpm). The *HindIII* restriction enzyme (cat. no. R0104; NEB) was added at a concentration of 1500 units per 7x10^6 cells and digestion was carried out overnight at 37°C with agitation (950 rpm). The restriction enzyme ends were filled in with Klenow (cat. no. M0210; NEB) for 75 minutes at 37°C using biotin-14-dATP (cat. no. 19524-016; Invitrogen), dCTP, dGTP, dTTP (cat. no. 10297018; Invitrogen) all at a final concentration of 30µM. This was followed by ligation with 50 units of T4 DNA ligase (cat. no. 15224-025; Invitrogen) for 4 hours at 16°C in a 5.5ml volume of ligation buffer (50mM Tris-HCl pH 7.5; 10mM $MgCl_2$; 1mM ATP; 10mM DTT; 100µg/ml BSA; 0.9% Triton X-100) per 7x10^6 cells. Subsequently the crosslink was reversed by overnight incubation at 65°C with the addition of 65µl of 10mg/ml per 7 million cells Proteinase K (cat. no. 03115879001; Roche), followed by an additional 65µl Proteinase K incubation for 2 hours at 65°C. An RNase treatment was performed using 15µl of 10mg/ml RNase A (cat. no. 10109169001; Roche) per 7x10^6 cells at 37°C for 1 hour. Two phenol/chloroform (P3803; Sigma) extractions were performed followed by DNA precipitation using 3M Sodium Acetate pH 5.2 (1/10 volume) and 100% Ethanol (2.5x volumes) overnight at -20°C. The DNA was spun down at 3200xg for 30 minutes 4°C and resuspended in 400µl TLE (10mM Tris-HCl pH 8; 0.1mM EDTA) and transferred to a new 1.5ml Eppendorf tube. Another phenol/chloroform extraction and DNA precipitation was performed with the DNA pellets being washed three times in 70% ethanol and quantified using the Quant-iT Pico Green kit (cat. no. P7589; Invitrogen).

To prevent the pull-down of non-ligated products the biotin from fragment ends was removed using the T4 DNA polymerase (cat. no. M0203; NEB) with 40µg of the Hi-C library DNA incubated for 4 hours at 20°C, followed by phenol/chloroform extraction and DNA precipitation overnight. The precipitated DNA was sonicated to a size peak of around 400bp using the Covaris E220 sonicator with the following settings: duty factor - 10%; peak incident power - 140W; cycles per burst - 200; time - 55 seconds. End repair was performed using T4 DNA polymerase (cat. no. M0203; NEB), T4 polynucleotide kinase (cat. no. M0201; NEB), Klenow (cat. no. M0210; NEB) in the presence of dNTPs in ligation buffer (cat. no. B0202; NEB) and incubated for 30 minutes at 22°C. DNA was purified using the QIAquick PCR purification kit (cat. no. 28104; Qiagen). A-tailing was performed with Klenow exo- (cat. no. M0212; NEB) and dATP incubated for 30 minutes at 37°C, followed by heat inactivation of the enzyme at 65°C for 20 minutes. A double-sided Agencourt AMPure XP bead (cat. no. A63881; Beckman Coulter) fragment size selection was performed. First, a 0.6x concentration of AMPure XP beads was used to remove large fragments. The supernatant was transferred to a new Eppendorf tube and the beads stuck to the magnetic separator (cat. no. 12321D; DynaMag-2 magnet; Invitrogen) were discarded. Additional AMPure XP beads were added to the supernatant to make up a 0.9x concentration of beads to DNA volume. After two washes with freshly prepared 70% ethanol, the DNA was eluted with 100µl of TLE (10mM Tris-HCl pH 8.0; 0.1mM EDTA). MyOne Streptavidin C1 Dynabeads (cat. no. 65001; Invitrogen) were used to isolate biotinylated ligation products in a binding buffer (5mM Tris pH 8, 0.5mM EDTA, 1 M NaCl) for 30 minutes at 22°C. While on the DynaMag-2 magnet, the C1 Dynabeads were washed twice with binding buffer and once with ligation buffer (cat. no. B0202; NEB). Illumina PE adapters were ligated onto Hi-C ligation products that were still bound to the streptavidin C1 beads using T4 DNA ligase (cat. no. M0202; NEB) with ligation buffer for 2 hours at 22°C slowly rotating. The streptavidin C1 beads were washed twice with wash buffer (5mM Tris, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween20) and once with binding buffer and finally resuspended in a final volume of 90µl NEBuffer 2. Hi-C DNA was amplified off the streptavidin C1 beads with 7 PCR cycles using the Illumina PE PCR 1.0 and 2.0 primers, followed by 0.9x AMPure XP beads purification. The concentration and fragment length profiles of the Hi-C libraries were determined with the 2100 Bioanalyzer (cat. no. G2939BA; Agilent) and libraries were pair-end sequenced on the Illumina HiSeq 2500 system.

2.14.2 Promoter Capture Hi-C (PCHi-C)

The capture of promoter containing *HindIII* fragments was carried out using the SureSelect XT Target enrichment (cat. no. 5190-4806, 5190-4831; Agilent Technologies) according to manufacturer's instructions as previously described (Schoenfelder et al., 2015). 37,608 biotinylated 120-mer custom RNA baits were designed by Simon Andrews as previously described (Mifsud et al., 2015), targeting the ends of 22,076 HindIII fragments overlapping Ensembl v75 (GRCh37) annotated promoters of

protein-coding, noncoding, antisense, snRNA miRNA and snoRNA transcripts. The targeting sequence needed to have a GC content ranging between 25% and 65%, containing no more than 2 consecutive Ns and be no more than 330bp from the restriction fragment ends. 500ng of the Hi-C library was used for the promoter capture, resuspended in 3.6µl H_2O along with Agilent Technologies hybridization blockers 1, 2 and a custom hybridization blocker (cat. no. 931108; Agilent Technologies). The Hi-C library DNA was denatured at 95°C for 5 minutes and allowed to hybridise in the presence of a hybridisation buffer with the RNA capture bait system at 65°C for 24 hours. 60µl of MyOne Streptavidin T1 Dynabeads (cat. no. 65601, Invitrogen) were washed three times in 200µl binding buffer, followed by incubation with the Hi-C DNA/RNA capture bait mixture in 200µl of binding buffer for 30min at 22°C. Streptavidin bound DNA was isolated using a magnetic separator (cat. no. 12321D; DynaMag-2 magnet; Invitrogen) followed by washes, 15 minutes in 500µl wash buffer I followed by 3 washes in 500µl wash buffer II at 65°C for 10 minutes. A final 300µl wash with NEBuffer 2 was performed and the beads were resuspended in a final volume of 30µl NEBuffer 2. A post-capture 4 cycle PCR was performed with Illumina PE PCR 1.0 and PE PCR 2.0 primers followed by a Agencourt AMPure XP bead (cat. no. A63881; Beckman Coulter) clean-up step. The concentration and fragment length profiles of the Hi-C libraries were determined with the 2100 Bioanalyzer (cat. no. G2939BA; Agilent) and libraries were pair-end sequenced on the Illumina HiSeq 2500 system.

2.15 ChIP-seq

ChIP-seq experiments were performed with Amanda J. Collier.

All buffers were pre-chilled to 4°C with cOmplete EDTA-free protease inhibitor (cat. no. 04693159001, Roche) freshly added. 15 million cells per ChIP were Accutase (cat. no. A1110501; Thermo Fisher Scientific) treated and collection in a 50ml conical tube, followed by a 300xg 5 minute spin at 4°C. Pellet was resuspended in PBS. Cells were double cross-linked with freshly prepared DSG (cat. no. 80424-50MG-F; Sigma) for 45 minutes at 22°C and subsequently with 1% methanol-free formaldehyde (cat. no. AGR1026, Agar Scientific) at a cell density of 10^8 cell in 45 ml media for 12.5 minutes at 22°C. Fixation was stopped with the addition of glycine at a final concentration of 125mM and incubated for 5min at 22°C. After two PBS washes, cells were resuspended in Wash buffer 1 (10mM Hepes pH 7.5; 10mM EDTA; 0.5mM EGTA; 0.75% Triton X-100) and incubated for 10 min at 4°C. After spinning at 3200xg for 5 minutes 4°C, nuclei were resuspended in 10ml Wash buffer 2 (10mM Hepes pH 7.5; 200mM NaCl; 1mM EDTA; 0.5mM EGTA) and incubated for 10 minutes at 4°C. Another 3200xg 5 minute at 4°C spin was performed followed by resuspension in 1ml of freshly made Lysis/sonication buffer (150mM NaCl; 25mM Tris pH 7.5; 5mM EDTA; 0.1% Triton X-100; 1% SDS; freshly dissolved 0.5% Sodium deoxycholate) per 12 million cells. Lysis was performed on ice for 30 minutes, followed by sonication 15 seconds on, 30 seconds off (Microson ultrasonic cell disruptor XL Misonix; output setting

4; 10-11 W) for 20 cycles. Optimal fragment size of 200-500bp is desired. Fragmented chromatin was spun down at 10000xg for 15min at 4°C and supernatant was transferred to a new tube and diluted 1:10 with ChIP dilution buffer (150 mM NaCl; 25mM Tris pH 7.5; 5mM EDTA; 1% Triton X-100; 0.1% SDS; 0.5% Sodium deoxycholate). 500µl was taken for the input and the remaining diluted supernatant was incubated with 5µg of antibody (see Table 2-19) overnight at 4°C. Magnetic protein A (120µl per IP) or protein G (180µl per IP) Dynabeads (cat. no. 10002D, 10004D; Invitrogen) were washed with Wash buffer A (50mM Tris pH 8; 150mM NaCl; 0.1% SDS; 0.5% Sodium deoxycholate; 1% NP40; 1mM EDTA) and blocked for 1 hour at 4°C with yeast tRNA (cat. no. AM7119; Invitrogen) and BSA (cat. no. B9000s; NEB). The pre-blocked beads were added to the antibody bound chromatin and incubated for 7-8 hours at 4°C. Subsequently the magnetic beads with the bound antibody chromatin complex were rinsed 1x with Wash buffer A, washed 2x with Wash buffer A, washed 1x with Wash buffer B (50mM Tris pH 8.0; 500mM NaCl; 0.1% SDS; 0.5% Sodium deoxycholate; 1% NP40; 1mM EDTA), washed 1x with Wash buffer C (50mM Tris pH 8; 250mM LiCl; 0.5% Sodium deoxycholate; 1% Igepal CA-630; 1mM EDTA) and rinsed with 1x TE buffer (10mM Tris pH 8; 1mM EDTA). Chromatin was eluted off the beads with 450µl of Elution buffer (1% SDS; 0.1M $NaHCO_3$ in H_2O). Additionally, 11µl Proteinase K (20mg/ml) and 5µl RNase A (10mg/ml) were added (including the input) and incubate at 37°C for 2 hours, followed by an overnight 65°C incubation to reverse the crosslink. DNA was purified using AMPure XP beads (cat. no. A63881; Beckman Coulter) and eluted in 40µl H_2O. DNA was quantified using the Qubit fluorometer (cat. no. Q33216; Invitrogen) dsDNA HS assay kit (cat. no. Q32851; Invitrogen). Libraries were prepared using the NEBNext Ultra II DNA library prep kit for Illumina (cat. no. E7645S; NEB) using the manufacturers protocol.

Table 2-19: List of antibodies used for ChIP-seq

Reactivity	**Catalogue number**
H3K4me1	ab8895; Abcam
IgG	315-005-003; Jackson Immune Research

2.16 Data processing and analysis

Table 2-20: Summary of datasets generated in lab and datasets used from publications, including MACS2 peak calling settings, number of significant peaks called, and effective genome size used with deeptools.

Histone mark and transcription factors	Cell type	Peaks	Tag size	Effective genome size (EGS)	EGS used in MACS2	Q value used in MACS2	Dataset origin
CTCF	Naïve	28017	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-7}$	(Ji et al., 2016)
H3K27ac	hPSC	34116	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-9}$	(Ji et al., 2016)
H3K27me3		224	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-9}$	(Theunissen et al., 2014)
H3K4me1		32998	75	2,747,877,777	$2.70x10^9$	$1.00x10^{-7}$	In lab
H3K4me3		21986	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-9}$	(Theunissen et al., 2014)
H3K9me3		6239	100	2,805,636,331	$2.70x10^9$	$1.00x10^{-9}$	(Theunissen et al., 2016)
OCT4		6481	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-9}$	(Ji et al., 2016)
CTCF	Primed	16223	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-7}$	(Ji et al., 2016)
H3K27ac	hPSC	19748	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-9}$	(Ji et al., 2016)
H3K27me3		5200	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-9}$	(Theunissen et al., 2014)
H3K4me1		34857	36	2,701,495,761	$2.70x10^9$	$1.00x10^{-7}$	(Rada-Iglesias et al., 2011)
H3K4me3		20832	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-9}$	(Theunissen et al., 2014)
H3K9me3		8219	100	2,805,636,331	$2.70x10^9$	$1.00x10^{-9}$	(Theunissen et al., 2016)
OCT4		7108	40	2,701,495,761	$2.70x10^9$	$1.00x10^{-9}$	(Ji et al., 2016)

2.16.1 Mapping and processing Hi-C and PCHi-C data - HiCUP

Hi-C data and PCHi-C data was mapped and artefacts were filtered out using HiCUP version 0.5.8-0.5.9 (Wingett et al., 2015) with the GCRh38 human genome build. HiCUP was run by Steven Wingett. The Hi-C protocol enriches for library fragments that contain two DNA fragments separated by a biotin incorporated restriction site. In order to align these chimeric reads (di-tags), HiCUP truncates reads at the restriction enzyme cut site, producing a contiguous genomic sequence, which should each map to a single restriction fragment. In addition, HiCUP performed filtering of experimental artefacts and uninformative di-tags. These may be a result of adjacent restriction fragments re-ligating, or ligation of a restriction fragment to itself resulting in circularisation. Furthermore, inefficient biotin removal from non-ligated DNA fragments results in a high proportion of DNA sequences without an overlapping restriction cut site, termed "internal fragments". HiCUP reports can be found in appendix.

2.16.2 Calling significant promoter interactions - CHiCAGO

Interaction significance was called using CHiCAGO version 0.2.4 (Cairns et al., 2016). CHiCAGO utilises a convolution of a negative binomial distribution and a Poisson distribution to create a background model. The model incorporates interactions arising from Brownian motion as a function of genomic distance along with random 'technical' noise arising from assay and sequencing artefacts. The assumption is that these two sources of background are independent and therefore can be combined into a Delaporte distribution.

Because of the high correlation of the two biological replicates, it was decided to merge them as part of the CHiCAGO pipeline. An interaction was considered significant if it had a score of 5 or more, based on previous empirical observations, which aimed to strike a balance between the saturation of regulatory features over promoter interacting regions vs random controls and the number of total interactions retained, which decreases with increased score threshold (Freire-Pritchett et al., 2017). For promoter regulatory region interaction analysis, an interaction with a score lower than 5 was retained if an equivalent significant interaction was present in the other cell type. For simplicity all network visualisation was done only with interaction score of 5 or more; however, network with a score of 3 or more were also constructed to validate cell specific interaction observations.

2.16.3 Chromatin state analysis

Chromatin state analysis was performed using ChromHMM (Ernst and Kellis, 2012, 2017) (hidden Markov model). Trim Galore quality trimmed and Bowtie2 (Langmead and Salzberg, 2012) aligned (GCRh38) BAM files were binarised using the `BinarizeBam` command with default 200bp bin settings. Both naïve and primed cell types were stacked to provide a single-genome annotation with the inclusion of ChIP-seq input samples as an additional feature. Model learning was performed on a range of states with 16 being chosen as the final number (Figure 2-4). These were further reduced into 7 states which were more biologically relevent (Active – H3K27ac, H3K4me3, H3K4me1; Polycomb repressed – H3K27me3; Bivalent – H3K27me3, H3K4me3, H3K4me1; Heterochromatin repressed – H3K9me3; Mixed - H3K27me3, H3K4me3, H3K4me1, H3K9me3; Background – low emission probability levels in all samples).

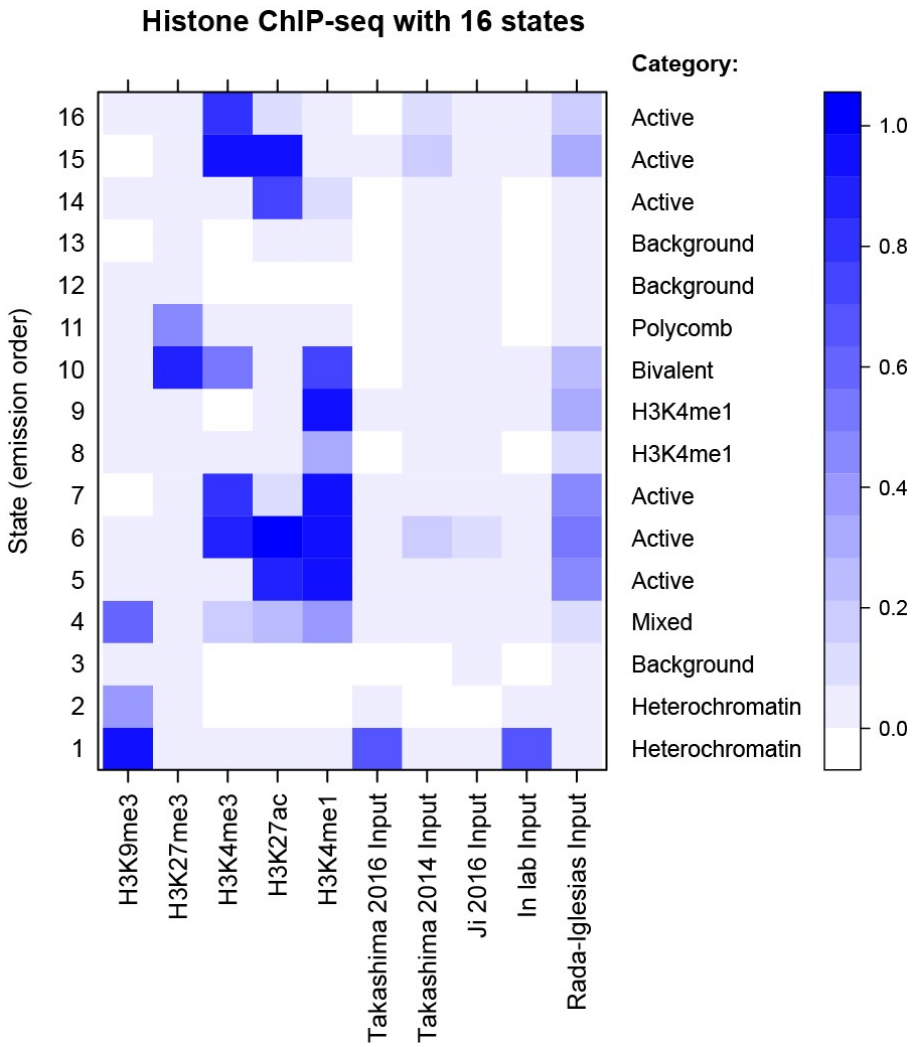


Figure 2-4: ChromHMM emission probability of 16 states for 5 histone modifications and their respective input sample. The 16 states were additionally reduced to 7 states based on the combination of histone modifications within each state: Active – H3K27ac, H3K4me3, H3K4me1; Polycomb – H3K27me3; Bivalent – H3K27me3, H3K4me3, H3K4me1; Heterochromatin – H3K9me3; Mixed - H3K27me3, H3K4me3, H3K4me1, H3K9me3; Background – low emission probability levels in all samples.

To assign a state to each *HindIII* fragment, the overlap of the 7 states with each restriction fragment was determined and reduced to a single final state based on the following rules: any single state superseded the background state; the bivalent state superseded the polycomb state; a mixture of multiple states was labelled as unclassified.

2.16.4 RNA-seq analysis

Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) quality trimmed and Hisat2 (Kim et al., 2015) aligned (GCRh38) BAM files (see Table 2-21) were loaded into Seqmonk (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk) along with Ensembl gene annotation release 85. Raw counts were exported for DESeq2 (Love et al., 2014) differential gene expression analysis, along with log2(FPKM) normalised values.

Table 2-21: RNA-seq dataset used.

Sample	Accession	Publication
Naïve H9 reset	ERR590398; ERR590400; ERR590399	(Takashima et al., 2014)
Primed H9	ERR590408; ERR590410; ERR590401	(Takashima et al., 2014)

2.16.5 ChIP-seq analysis

Peak calling on Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) quality trimmed and Bowtie2 (Langmead and Salzberg, 2012) aligned (GCRh38) BAM files was performed with MACS2 (Zhang et al., 2008). The settings used and number of peaks called is detailed in Table 2-20.

2.16.6 Genome browser tracks

Normalised bigwig files for genome browser visualisation were produced using Deeptools (Ramírez et al., 2016). For ChIP-seq samples the BAM files were normalised with the reads per genomic content (RPGC) method, ignoring chrY and chrMT. This essentially divides the number of reads per bin with a scaling factor (inverse of the sequencing depth of the sample) for 1x average coverage. The effective genome size used for each sample is listed in Table 2-20. A 10bp bin size with a 200bp read extension was chosen with a bigwig file output. Finally, the inputs were subtracted from the sample. For RNA-seq tracks, the BAM files were normalised using the DESeq2 scaling factor (Naïve - 0.34576529; Primed - 1.873937595), with a default bin size of 50bp. As the RNA-seq was directional, separate files were produced for each strand. Tracks were visualised using the WashU Epigenome Browser v46.1 (Zhou et al., 2011, 2013, 2015).

2.16.7 Hi-C analysis and the definition of A/B compartments and TADs.

The Hi-C analysis was performed by Csilla Varnai using HOMER v4.7 (http://homer.salk.edu/homer/) (Heinz et al., 2010). Interaction matrices were binned at 25kb and 250kb resolutions and corrected using the iterative correction algorithm (Imakaev et al., 2012). A correlation matrix was constructed from normalised data at 250kb resolution and principle component analysis was performed to call A/B compartments(Lieberman-Aiden et al., 2009). The assignment to A or B compartment was based on the overlap with H3K4me3 ChIP-seq peaks (see Table 2-20). TADs were called using directionality indices (Dixon et al., 2012) of interactions 1Mb upstream and downstream from the centre of a 25kb sliding window with a 5kb step interval. The directionality indices were smoothed by calculating a running average over a 25kb window and pairs of local maxima and minima with a standard score

difference (TAD ΔZ score) above 2 constituted a TAD call. Finally, the TAD ends were extended outwards to the bins with no directionality bias.

2.16.8 Promoter capture Hi-C (PCHiC) analysis

Analysis of significantly called PCHiC interactions was performed using custom scripts in R and Python. Interactions networks were constructed using the R igraph v1.2.1 package (Csardi and Nepusz, 2006). Community detection was performed using the multi-level optimisation algorithm (Blondel et al., 2008) implemented within igraph. For a subnetwork to be split into individual communities a modularity score of 0.7 or more was required. Network visualisation was performed within Gephi v0.9.2 (Bastian et al., 2009) using the ForceAtlas2 layout algorithm (Jacomy et al., 2014). The multi-dimensional scaling (MDS) layout within Gephi, developed by Wouter Spekkink (http://www.wouterspekkink.org), was used to obtain distance representative layouts of individual subnetworks. Genome tracks were visualised with either the KaryoploteR v1.6.3 R package (Gel et al., 2017) or using the WashU Epigenome Browser version 46.1 (Zhou et al., 2011, 2013, 2015). Circular ideograms of the Hox loci were created using the Circos visualisation tool (Krzywinski et al., 2009).

3 B cell development and epigenetic mechanisms underpinning recombination in a human immunoglobulin transgene model

3.1 Background

Numerous studies have sought to uncover the complex roles of the Igh and Igl (immunoglobulin light chain) in B cell development. The Igl recombines and is expressed at the small pre-B developmental stage (fraction D), which is followed by the formation of the BCR through the pairing of the light chain (LC) and the heavy chain (HC) proteins. By functionally inactivating both endogenous immunoglobulin Igl loci (Igλ and Igκ; $Igl^{-/-}$) in the mouse, B cell development is arrested at the pre-B developmental stage, unable to progress without the BCR (Zou et al., 1995, 2003). The reciprocal experiments were also performed in the Igh locus, where the targeted deletion of the constant region genes functionally inactivated the locus. By only targeting the heavy chain constant domain genes (C_H), developmentally important genes interspersed in the intergenic regions of the Igh, such as *Adam6,* remained functionally intact (Featherstone et al., 2010; Han et al., 2009). The inactivation of the Igh resulted in a developmental block at the pre-B-I cell stage (fraction BC), where the absence of the HC prevented the formation of the pre-BCR that is required for developmental progression (Ren et al., 2004). A later study, also from Marianne Brüggemann's lab, demonstrated that in $Igl^{-/-}$ mice the high selection pressures of B cells without a LC can result in the spontaneous production of HCAbs, resulting from random mutation within the C_H1 domain (Brüggemann et al., 2010; Zou et al., 2007). Additional studies performed using camel (dromedary) HCAb recombined sequence have shown that HCAbs can be successfully produced in mice and are capable of driving mouse B cell development (Zou et al., 2005). The ability of VHHs and mouse variable fragments to drive B cell development expands our ability to utilise HCAbs for potential therapeutic applications. Additionally, by utilising humanised mice, the longer and broader distribution of amino acids within the human CDR_H3s could potentially increase the intrinsic ability of HCAbs to recognise conventionally inaccessible epitopes (Lee et al., 2014; Muyldermans, 2013), in addition to circumventing the production of HAMAs.

3.1.1 Crescendo Biologics and the Crescendo Mouse

The biopharmaceutical company Crescendo Biologics was created to harness the therapeutic potential of fully human heavy chain only fragments. Their successful creation of the Crescendo mouse platform gave them the ability to harness the mouse *in vivo* antibody maturation process and obtain high affinity human HCAbs, termed Humabody VH, against desired antigens with exceptional biophysical characteristics. Currently, Crescendo Biologics is pursuing the development of multi-functional Humabody therapeutics in oncology, utilising Humabody VH to selectively activate tumour-targeted T cell engagers within the tumour microenvironment, which would avoid the toxicity of current chimeric antigen receptor (CAR) T cells therapies (Maus and June, 2016).

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-1

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-2

The following text has been redacted due to sensitivity of the material

3.1.2 Predicted B cell development in CTG mice

From our current knowledge of the CTG2 mouse, it is possible to draw parallels with other models of B cell development. Without the C_H1 domain, it is reasonable to postulate that the surrogate light chain (SLC) would be unable to bind the HC in CTG2 mice in the same capacity as in WT. The SLC pairing with the HC and the formation of the pre-BCR constitutes an important B cell developmental checkpoint. As a result, the SLC knockout models are a valuable resource for our understanding of B cell development in CTG2 mice.

3.1.2.1 SLC knockout models display an alteration in B cell development

The invariant SLC is encoded by the *VpreB1/2* and *λ5* genes and shares structural homology to the LC. In SLC triple knockout ($SLC^{-/-}$; $VpreB1^{-/-}$, $VpreB2^{-/-}$, $\lambda5^{-/-}$) mice, the observation that the pairing of two HCs results in allelic exclusion, while non-pairing results in recombination of the second allele, indicates that pre-BCR signalling is at least partially intact (Shimizu et al., 2002). Studies with truncated HCs revealed that the SLC is not essential for pre-BCR signalling and function (Shaffer and Schlissel, 1997). The transmembrane signal transduction proteins Igα and Igβ have to be intact for developmental progression. So, while a defective HC leads to developmental arrest at the pro-B stage (fraction BC), a defective SLC results in decreased proliferation, but allows B cells to progress to the pre-B stage (fraction C') and beyond (Shaffer and Schlissel, 1997; Shimizu et al., 2002). The $SLC^{-/-}$ has also been shown to impact on the total number of B cells in the spleen (Shimizu et al., 2002). In pre-B cells, most heavy-chains associated with the SLC are found in the ER, whereas most HCs associated with the LC at the immature stage of development are found on the plasma membrane. A study using localisation sequences attached to the HCs to redirect them into the *trans*-Golgi network (TGN), demonstrated that the pre-BCR is capable of signalling and inducing pre-BCR dependent events from other cellular compartments other than the plasma membrane (Guloglu and Roman, 2006). The ability to signal from other compartments is supported by previous studies that demonstrated the ligand-independent (autonomous) signalling capability of the pre-BCR, which has been proposed to result from the interaction of multiple pre-BCR receptors through their charged domains on the SLCs (Köhler et al., 2008; Ohnishi and Melchers, 2003; Winkler and Mårtensson, 2018).

Independently of the HC, crosslinking of Igβ has been implicated in transducing signals that trigger developmental progression from the pro-B stage (fraction BC) (Nagata et al., 1997). Igα/Igβ has been shown to have independent, but critical functions in B cell developmental progression (Papavasiliou et al., 1995; Reichlin et al., 2001). The HC in the absence of the SLC/LC relocates from the ER to the plasma membrane together with Igα (Su et al., 2003). Additionally, truncated HCs in $LC^{-/-}$ mice are deposited on the cell surface together with Igβ (Zou et al., 2008). Interestingly, a mouse strain whose B cell development was reliant on IgG1 instead of IgM showed a decreased dependence on Igα for maintenance (Waisman et al., 2007). The levels of IgM in serum of $SLC^{-/-}$ mice were not significantly affected, although they seemed to have slightly reduced immune response to antigens (Shimizu et al., 2002).

3.1.2.2 B cell development with a IgG-BCR and the importance of secreted IgM (sIgM)

The importance of the IgM class is highlighted by its presence in all vertebrates (Flajnik, 2002). The μ and δ are the only heavy chains lacking a cytoplasmic tail and as a result are thought to be more reliant on the Igα/β (CD79) transmembrane proteins. Igα/β associate with the pre-BCR/BCR and transduce signals through their cytoplasmic ITAM modules (Geisberger et al., 2003). Igα/β have been shown to associate with all classes of immunoglobulins. They are required for the surface expression of IgM, IgA, but not IgG2b. In IgG-BCR mouse models, the use of the γ constant region (forming IgG) instead of μ (forming IgM) was shown to impact B cell development (Waisman et al., 2007). A reduction in the transitional B cells and an increase in MZ B cells was observed in the spleen. In addition, there was a higher number of fraction A-C B cells with severely diminished numbers of fraction D and extending to fraction E B cells, suggesting a developmental block at the pre-BCR stage of development (Waisman et al., 2007). The IgG-BCR mouse models also demonstrated the dispensability of Igα for IgG-BCRs but not that of Igβ. The truncation of both Igα/β fully blocked B cell development from the pro-B (fraction B) stage of development (Song et al., 2016; Waisman et al., 2007),

The following text has been redacted due to sensitivity of the material

It is known that IgG has a more profound response to antigen stimulation compared to IgM based on its ability to induce higher clonal expansion and production of antibodies that results from its unique signalling (Martin and Goodnow, 2002). The difference between IgM and IgG signalling was shown to be due to the inhibitory action of CD22 on IgM and IgD BCRs. The IgG BCR is able to circumvent this inhibition through its unique cytoplasmic tail that prevents phosphorylation and therefore activation of CD22 (Wakabayashi et al., 2002). This suggests that the use of an IgG in early B cell development and the altered B cell development observed in mouse models could be a consequence of the unique signalling of the IgG receptor.

The following text has been redacted due to sensitivity of the material

The importance of IgM in B cell development extends beyond its signalling capacity as a BCR. Natural IgMs or secreted (sIgM) (produced even under sterile conditions) are polyreactive and have been described to recognise self-antigens (Ferry et al., 2007). They are produced in the bone marrow and the spleen at high levels by $CD5^-$ plasma cells and B-1 cells, which typically reside in the peritoneal cavity (Choi et al., 2012; Reynolds et al., 2015; Savage and Baumgarth, 2015). Besides their protective roles during infection, they have also been implicated in cellular debris clearance and the reduction of autoantigen levels (Ehrenstein and Notley, 2010). The deletion of the secretory sequence in exon 4 of the Cμ gene that ablates sIgM (Ehrenstein et al., 1998) has also been shown to impact normal B cell development.

The impact is most noticeable on the peripheral B cell populations, such as in the spleen, where a significant increase in MZ B cell and a significant decrease in FO B cells was observed (Baker and Ehrenstein, 2002; Nguyen et al., 2015; Tsiantoulas et al., 2017). One suggested mechanism for this change is the increased BCR signalling due to the failed clearance of antigens by the sIgM, which would direct B cells towards the MZ cell fate (Tsiantoulas et al., 2017). However, contradictory studies done in BCR transgene mice, and Btk or Aiolos knockout mice, suggest weaker BCR signalling directs cell fate towards MZ B cells and stronger towards FO B cells instead (Casola et al., 2004; Pillai and Cariappa, 2009; Pillai et al., 2004). Yet, even within these studies one BCR transgene model specific for a low-dose self-antigen resulted in a contrary development of B cells into the MZ fate (Wen et al., 2005). In addition, the observed increase in B-1 B cell population upon sIgM ablation suggest that MZ and B-1 cells respond to comparable levels of BCR signalling (Boes et al., 1998), which for B-1 cells has been shown to be strong rather than weak BCR signalling (Berland and Wortis, 2002). Another contrary result, attributed to a previously unidentified anergic population, was again generated in a later study which showed a decrease in the peritoneum B-1 B cell population, while maintaining normal number in the spleen (Nguyen et al., 2015). Reconciling these findings will be important in elucidating the exact mechanism driving cell fate of peripheral B cells. One possibility could be that the self-recognition and foreign antigen recognition give rise to different levels of BCR signal response that could potentially be quantified by monitoring the levels of calcium flux post BCR stimulation.

Beyond peripheral B cells, the impact of ablating sIgM extends all the way back to the bone marrow. An increase in fraction A/B with a decrease in fractions D/E/F was observed in the bone marrow of $sIgM^{-/-}$ mice (Nguyen et al., 2015). In addition, an increase in the generation of autoantibodies was

observed (Boes et al., 2000; Ehrenstein et al., 2000) and attributed to the breakdown of central tolerance in the absence of sIgM$^{-/-}$, which resulted in an altered B cell antibody repertoire (Nguyen et al., 2015). Interestingly, the breakdown of the central tolerance checkpoint at fraction E, which results in higher autoantibodies in the periphery, did not result in an increased number of fraction E B cells. This suggests that rather than more autoreactive BCRs bypassing central tolerance, it is possible that foreign antigen BCRs are negatively selected out due to the lack of proper stimulatory signalling from the sIgM. The receptor for the Fc portion of IgM (FcμR), restricted to B cells in mice and lymphocytes (B, T and natural killer (NK) cells) in humans (Kubagawa et al., 2017), is the most obvious target of sIgM. A knockout of FcμR shared a similar phenotype with the sIgM$^{-/-}$ models, displaying a decreased number of fraction D/E B cells and an increase in peritoneum B-1 B cells (Choi et al., 2013b). Additionally, the FcμR was shown to co-localise with the intercellular IgM and control tonic signalling through the internalisation of surface BCRs of the immature B cell subset (fraction E) (Nguyen et al., 2017). These studies demonstrate an interplay between FcμR and the IgM either in the secreted or membrane bound form. Yet, there have been four different FcμR$^{-/-}$ mouse models generated by independent labs that each display very different phenotypes, most likely from the different exon combinations targeted in each study (Nguyen et al., 2017; Wang et al., 2016). One certain commonality to all knockouts is the production of autoantibodies that suggest an alteration in BCR signalling that bypasses central tolerance, but are insufficient to cause pathology associated with autoimmune diseases (Wang et al., 2016).

3.1.2.3 The pre-BCR checkpoint and autoreactive antibodies

The following text has been redacted due to sensitivity of the material

The pre-BCR is present on the cell surface of fraction BC B cells at low levels. The HCs innate ability to moderately aggregate, forming multimeric complexes capable of signalling is enhanced by SLC mediated aggregation (Meixlsperger et al., 2007). The non-Ig tail of λ5 has been demonstrated to play a vital role in pre-BCR signalling and surface levels (Ohnishi and Melchers, 2003), despite having functional allelic exclusion in λ5$^{-/-}$ and SLC$^{-/-}$ mice (Boekel et al., 1998; Shimizu et al., 2002). In the absence of the SLC, lower levels of signalling may be compensated by HCs with a higher propensity for self-recognition and aggregation. Interestingly, positively charged arginine residues, characteristic of anti-DNA anti-nuclear antibodies (ANAs)(Radic et al., 1993), are found in the λ5-tail (von Boehmer and Melchers, 2010). The frequency of ANAs is increased in the absence of the SLC, with CDR_H3 basic residues unusually originating from D_H gene segments. Despite an increase in ANAs, no pathology was

observed in the $SLC^{-/-}$ mouse. An alternative route to the generation of ANAs is via somatic hypermutation (SHM). This was demonstrated when a SHM sequence was reversed to its original germline amino acid (aa) configuration and resulted in the abolishment of auto-reactivity (Keenan et al., 2008; Schroeder et al., 2012).

A skew towards MZ B cells in the numbers of splenic B cells was also observed in the $SLC^{-/-}$ mice, resulting from a decrease in FO B cell numbers rather than an increase in MZ B cells (Keenan et al., 2008; Ren et al., 2015). A number of studies have suggested the idea that BCR signalling strength drives lineage progression into peripheral B cell subpopulations (Casola et al., 2004; Heltemes and Manser, 2002; Pillai and Cariappa, 2009; Pillai et al., 2004; Tsiantoulas et al., 2017; Wen et al., 2005). In one such study, high levels of BCRs on the cell surface during normal development have been linked to negative selection, resulting from increased signalling that is characteristic of auto-reactive BCRs. However, modest levels of surface BCRs, obtained from alterations in transgene copy numbers, resulted in higher B cell numbers in the MZ-like cells (Heltemes and Manser, 2002). Additionally, a link between a subset of the V_H gene repertoire, auto-reactivity and the B-1 B cells has been made (Carmack et al., 1990; Lam and Rajewsky, 1999). For example, B cells with self-reactive specificity for phosphatidylcholine (PtC) have been shown to segregate into the B-1 subset, while B cells that do not recognise PtC, segregate into the B-2 subset (Mercolino et al., 1989; Arnold et al., 1994). B-1 B cells share many similarities with MZ B cells such as a bias towards a self-recognizing BCR repertoire (Paul, 2012).

Interestingly, the polyreactive chimeric receptors (pre-BCR like) from the Köhler *et. al.* study were also expressed at much lower levels on the cell surface compared to the non-autoreactive receptors (mature BCR like) (Köhler et al., 2008). A similar observation was made in $SLC^{-/-}$ mice, where lower levels of surface IgM were present on immature B cells compared to WT (Ren et al., 2015). The lower levels may result from higher abundance of polyreactive BCRs that passed through the pre-BCR negative selection. By depleting MZ B cells and monitoring autoantibody production, it was suggested that IgG ANAs are primarily produced by $CD21^{-}23^{-}$ B cells (B-1 like cells (Keenan et al., 2008)), whereas anti-dsDNA IgMs were produced by MZ B cells (Ren et al., 2015). The same group also reported that autoreactivate B cells in $SLC^{-/-}$ are also increased in FO B cells (Grimsholm et al., 2015).

SLC plays a major role in the repertoire selection at the pre-BCR checkpoint, namely the V_H domains that are unable to pair with the SLC and display autoreactivity are selected out (V_H81X, V_HQ52 and nearly half of the V_HJ558 domains). In $\lambda5^{-/-}$ mice, this negative selection is non-functional (Boekel et al., 1997). As a result, these V genes were observed in peripheral B cells, namely FO B cells (Grimsholm

et al., 2015). Interestingly, despite the breakdown of central tolerance in the bone marrow, peripheral tolerance was still intact and B cells entering the memory B cell and plasma cell pools did not contain V genes defective in pairing. Whether this is also true in B cell development reliant on IgG1 instead of IgM is unknown (Waisman et al., 2007).

3.1.2.4 Negative selection of auto-reactive µHC in humans - signalling without the SLC

Based on a study done in a single patient with a particular λ5 mutation, HCs deposited on the plasma membrane in human B cells are unable to signal activation and developmental progression. Instead they undergo apoptosis (Minegishi et al., 1998). This opens up the possibility that the germline HCs of humans do not have the same level of autoreactivity as those in mice and ANAs are rather derived from SHM (Schroeder et al., 2012). The strength of the pre-BCR signalling is the determinant for negative selection (Conley and Burrows, 2010). The possibility that the mutation resulting in λ5 deficiency causes a rare phenotype unlike the one studied in mice cannot be ruled out. An analysis of human HC transcripts revealed that about 9% will produce proteins that can be expressed on the cell surface without the SLC; however, their characteristically autoreactive CDR_H3 lead to B cell apoptosis (Minegishi and Conley, 2001). Whether the signalling mechanism and thresholds function differently in humans is debatable.

3.1.3

The following text has been redacted due to sensitivity of the material

3.1.4 Russell body formation

The following text has been redacted due to sensitivity of the material

Russell bodies (RBs) are distended cisternae of the endoplasmic reticulum (ER) filled with immunoglobulin aggregates. Plasma cells with RB formation and a defect in immunoglobulin secretion are called Mott cells (Alanen et al., 1985).

Transfecting a truncated heavy chain plasmid construct (lacking C_H1 domain) into a healthy myeloma cell line, which produces and secretes λ light chain, results in the formation of RBs. In a cell line that does not produce λ light chains, no substantial RBs were detected. The presence of RB does not seem to affect cell division (Valetti et al., 1991). Later work showed that RB formation can occur both in the presence and absence of the immunoglobulin light chain. When overexpressing the truncated heavy chain, the ER-associated degradation (ERAD) is overwhelmed and leads to RB formation (Mattioli et al., 2006). RB structures formed in the presence of the light chain (rough RB) differ from ones formed in their absence (smooth RB). They differ not only by the presence of ribosomes on the surface, but also by the structure of the aggregates within RBs (Mattioli et al., 2006). The ability of the LC to bind an HC lacking the C_H1 domain has been demonstrated numerous times in the past (Hendershot et al., 1987; Valetti et al., 1991). With Ac38 staining it was shown that despite the ablation of the C_H1 domain, the light chain still interacted with the heavy chain via the V domain. This leads to aggregation through the C_L that would normally interact with the C_H1 domain (Mattioli et al., 2006). RB formation resulting from LC dimer formation (Kaloff and Haas, 1995; Leitzgen et al., 1997; Mattioli et al., 2006) is somewhat counterintuitive as usually the light chain has a BCR anti-aggregation role that prevents strong autonomous signalling (Meixlsperger et al., 2007). LCs are important not only in development, but also in facilitating the correct folding of HCs (Lee et al., 1999). In $LC^{-/-}$ mice, the aggregation of full-length HCs in the ER leads to HC toxicity and cell death (Feige et al., 2009; Corcos et al., 2010).

HCs without the C_H1 domain are secreted in cell lines (Hendershot et al., 1987), *in vivo* models (Geraldes et al., 2007; Zou et al., 2007) and in Camelids that naturally produce HCAbs. Secretion of HCs was not observed in cell line models for RB (Valetti et al., 1991) which suggests certain types of HCs have a higher tendency to aggregate (Hasegawa et al., 2014). Alternatively, perhaps analogous to

a subset of light chains and the requirements of HCs to escape BiP, they fail to assemble and form dimers that are a prerequisite for secretion (Kaloff and Haas, 1995; Leitzgen et al., 1997). A study of functional IgGs and their propensity to form RB suggests the variable domains of the HC and LC, encoding physiochemical properties, govern the severity of RB formation. They observed that RB-forming IgGs deteriorated cell health and viability, which speculatively suggest HCs prone to RB formation are under negative selection in natural sources of HCAbs where RB formation has not been reported (Stoops et al., 2012). The formation of RB seems to be an additional quality control checkpoint before the release of immunoglobulins. By storing aggregation prone immunoglobulins within the ER the plasma cells prevent potentially harmful accumulation of extracellular immunoglobulins (Hasegawa et al., 2014).

3.1.5 Hypothesis

The following text has been redacted due to sensitivity of the material

3.1.6 Aims

The following text has been redacted due to sensitivity of the material

3.2 Reduced fraction D and appearance of an intermediate fraction E F population in CTG mouse bone marrow

I started the project by performing a comprehensive characterisation of the different B cell developmental fractions (Hardy et al., 1991) (Figure 1-2) in CTG2 mice with the aim of comparing them to the WT counterpart (Figure 1-1). I initially used a well-established sorting panel that was used for a number of past projects within the lab (Bolland et al., 2016; Koohy et al., 2018). The panel used the CD25 marker to identify fraction C' and D. However, the detection of the CD25 marker proved challenging in CTG2 mice and led me to develop a panel with alternative markers that included CD24 and AA4.1 (CD93).

Examination of the numbers of total cell extracted from the bone marrow of CTG2 and WT mice revealed very little difference. On average 128 million cells were extracted from CTG2 mice (n=8) and 115 million cells from WT (n=5). The only noticeable difference was the higher standard deviation of 47 million in CTG2 cell counts compared to 9 million in WT. I extended this analysis to the total counts of the $B220^{+}$ population from flow cytometry data in the bone marrow, which proved to be equivalent between CTG2 and WT (Figure 3-3 a). In spleen, the overall $B220^{+}$ counts are lower in CTG2 mice than WT; however, they do not reach statistical significance and comparing the weight of WT and CTG2 spleens also reveals no differences (Figure 3-3 b). This suggest that the B cell compartment is intact in CTG2 mice.

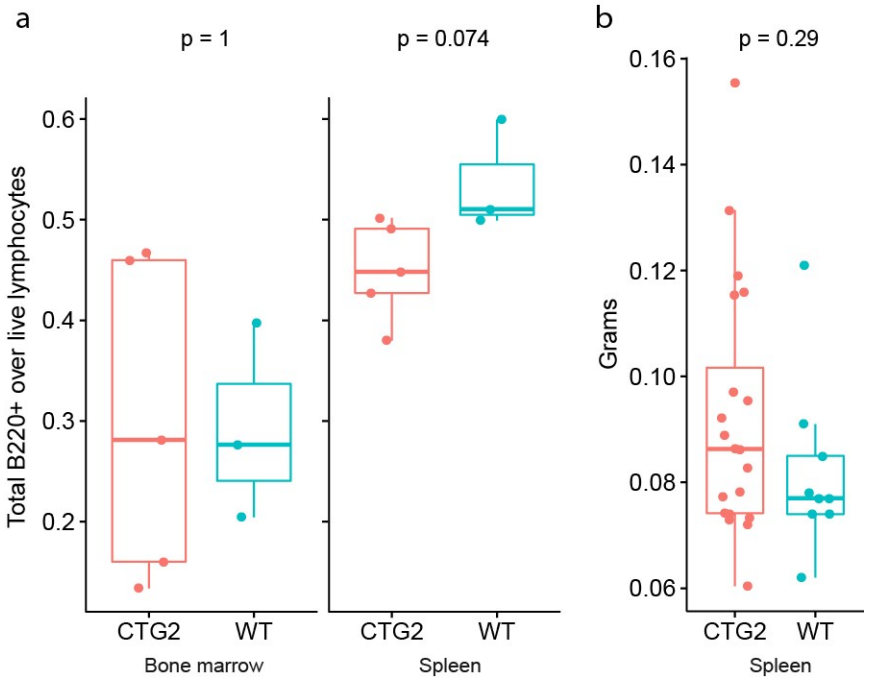


Figure 3-3: CTG2 mice show no differences in total number of B220 cells or size of spleen compared to WT mice. (a) Quantification of total B220 positive cells over live lymphocytes in the bone marrow and in the spleen of WT (n=3) and CTG2 (n=5) mice. P-values were calculated with an unpaired two-sample Mann-Whitney/Wilcoxon rank sum test with continuity correction. (b) Whole spleens weights from WT (n=9) and CTG2 (n=20) mice.

Examination of flow cytometry plots with the B220 and CD19 B cell markers revealed differences in the profile of the B220 high and low population between the CTG2 and WT mice (Figure 3-4). The B220 low population appears to be reduced in CTG2 mice, while the B220 high population displays a much broader spread of intensity values along the CD19 axis. These initial observations suggest there are differences in B cell developmental subsets as would be expected from published knockout models (Brüggemann et al., 2006; Nguyen et al., 2015; Shimizu et al., 2002; Waisman et al., 2007). In addition, the B220 high population in CTG2 seems to maintain lower intensity values compared to WT.

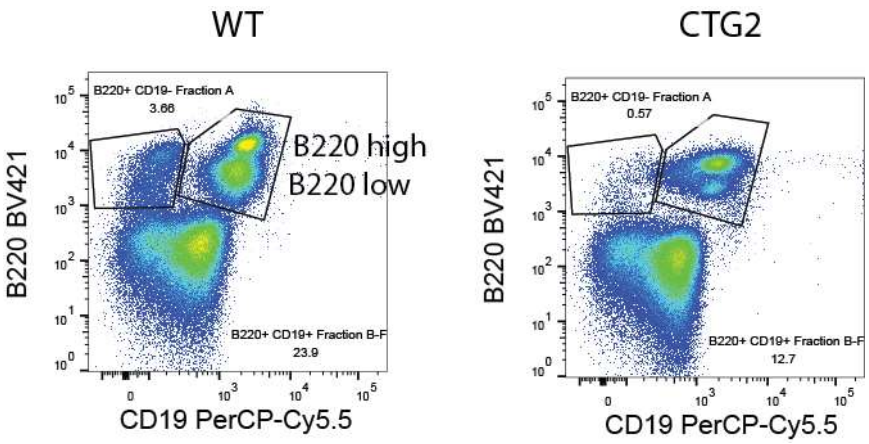


Figure 3-4: Differences in B cell bone marrow composition between WT and CTG2 mice.
B220 vs CD19 pseudo-coloured dot plots show an altered profile of the B220 low population in CTG2 mice compared to WT. CTG2 B220 high B cells have a greater spread along the CD19 axis compared to the B220 high CD19 high WT population (n>10). In addition, a less distinct population of B220 positive CD19 negative B cells is seen in CTG2 mice.

Due to the high dimensionality of flow cytometry data (6 parameters measured) and the possible unknown alterations in B cell developmental progression, I decided to use analysis approaches popularised and developed by the mass cytometry field, such as SPADE (Bendall et al., 2011; Qiu et al., 2011) and tSNE (Maaten and Hinton, 2008). Classic analysis allows two parameter comparisons such as those shown in Figure 3-4, meaning if a flow cytometry panel contains 6 parameters, 15 individual pairwise comparison plots would be required to have an overview of the data. The gated populations would require additional plots, making such analysis intractable for an unbiased examination of the data. I chose to use FlowSOM, which utilises self-organising maps (SOMs) to group cells into a predefined number of clusters based on their cell surface marker expression similarity. The final SOM is visualised as a minimal spanning tree (MST), allowing for example parameter and cluster composition overlays. The final visualisation is equivalent to plots produced with SPADE, with the distinct advantage of being much less computationally taxing and allowing the use of all captured events in the analysis.

From conventional pairwise pseudo-coloured dot plots it was clear that there were certain populations present in CTG2 mice which do not perfectly fall into the predefined B cell developmental

fractions. One such population is the intermediate population between fraction E and F (Figure 3-5 a). By plotting CD24 vs AA4.1 it is possible to view the transition from fraction BC all the way to fraction F. One striking difference between WT and CTG2 mice is the size of the $AA4.1^{+}$ and CD24 high population that correspond to fractions C' D and E. WT mice have a much larger fraction D and E population compared to CTG2 mice. The individual FlowSOM MSTs for WT and CTG2 capture the developmental progression along with the noted differences (Figure 3-5 b). The decreased proportion of fraction D and E cells is clearly shown, while an increase in fraction BC in CTG2 mice seems to be present. A dominant feature in CTG2 mice is the intermediate EF (intEF) population expansion, highlighted by the CD24 AA4.1 dot plots (Figure 3-5 a). More interestingly, a large portion of ungated cells is revealed not only clustering together with intermediate EF cells, but also forming a prominent feature upstream of intEF cells (Figure 3-5 b). This population, not captured by my manual gating strategy, very nicely illustrated the power of a global unbiased visualisation of flow cytometry data.

The manual gating of WT cells accounts for a large proportion of all the B220 positive cells within the FlowSOM analysis. However, in CTG2 mice there are large populations of ungated cells. In order to uncover what these may be, I overlaid the expression intensity of each measured cell surface marker onto the MST (Figure 3-6). One important aspect to note when interpreting the expression overlay plots is the range of values. Because the cells used for the analysis were all B220 positive, the low expression values for this marker does not mean that B220 is not expressed, as opposed to the other markers. Examining the intermediate EF clusters, which have not been fully captured by my manual gates, shows that while they have downregulated AA4.1 as expected for fraction F cells, their CD24 levels remain high, which is characteristic of fraction E. With the assumption that an IgG BCR mirrors the phenotypes produced by an IgM BCR, it is interesting that only a subset of these cells are expressing surface IgG that constitute my manually gated fraction E population. This suggests that the intermediate population contains cells that have failed to produce a functional cell surface expressed BCR. Normally such cells would undergo apoptosis after multiple failed attempts at V gene replacement, but instead the cells accumulate. The next set of cell clusters upstream of the intermediate EF gated population has not been captured by my gating strategy (Figure 3-5 b). These cells have downregulated AA4.1 and CD24, but their IgG expression remains low. This suggests that even though the cells now resemble fraction F more closely, they still do not have a functional cell surface BCR that would allow them to move forward, much like the intEF cells.

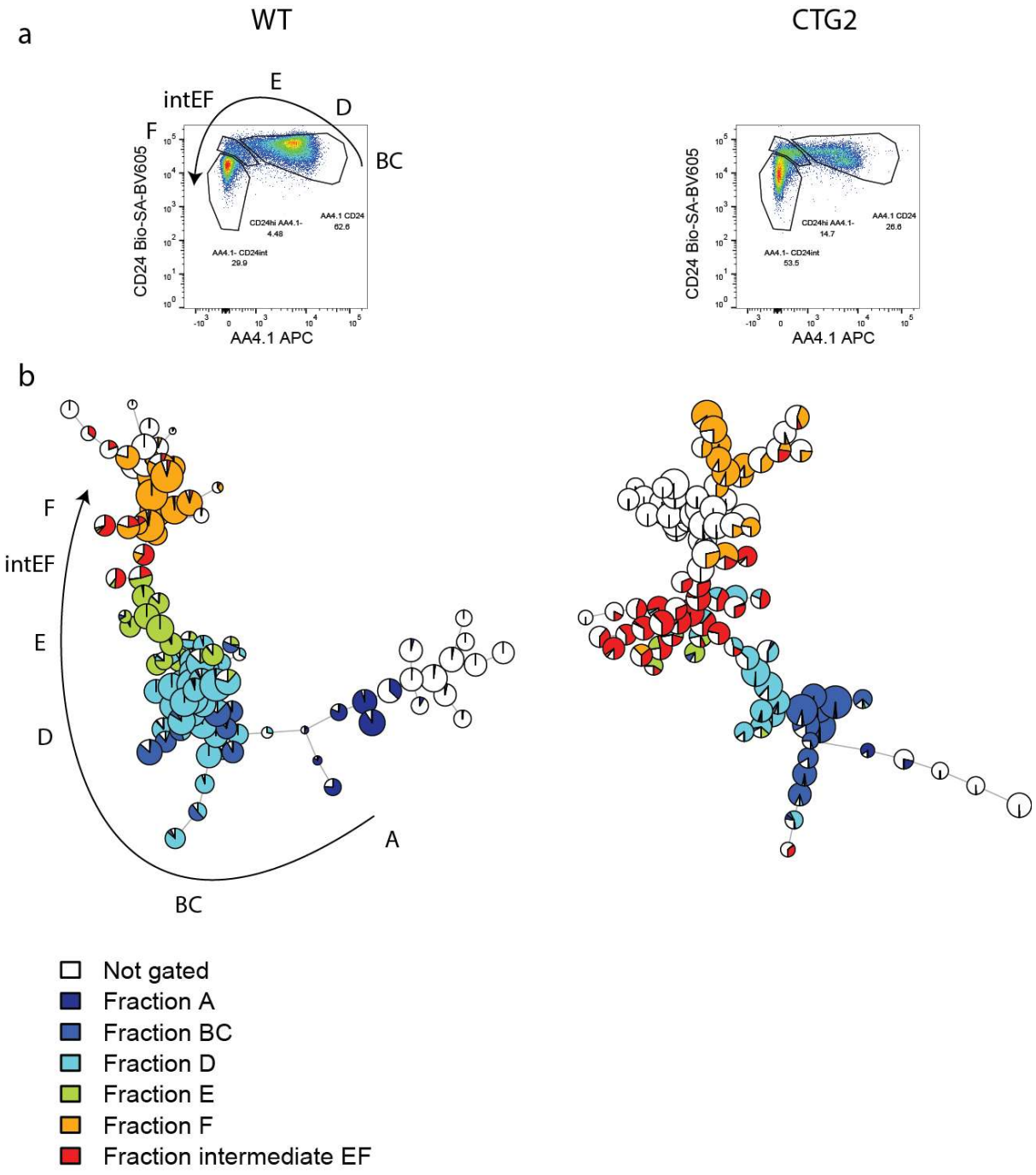


Figure 3-5: Flow cytometry analysis of B cell developmental fractions in WT and CTG2 mice reveals an intermediate population.
(a) CD24 vs AA4.1 pseudo-coloured dot plots highlighting an intermediate population between fractions E and F (intEF) in CTG2 mice. CD24 vs AA4.1 captures fractions BC, D, E and F on a single plot allowing a simple comparison between WT and CTG2. (b) Separate FlowSOM minimal spanning trees of WT and CTG2 flow cytometry data with manually gated fractions overlaid over the nodes. Each node represents a cluster of cells, with each pie chart giving the proportions of manually gated cells within that cluster. The size of the node is proportional to the cell numbers present within each cluster. Each MST was built from cells gated on lymphocyte, singlet and B220^{+}.

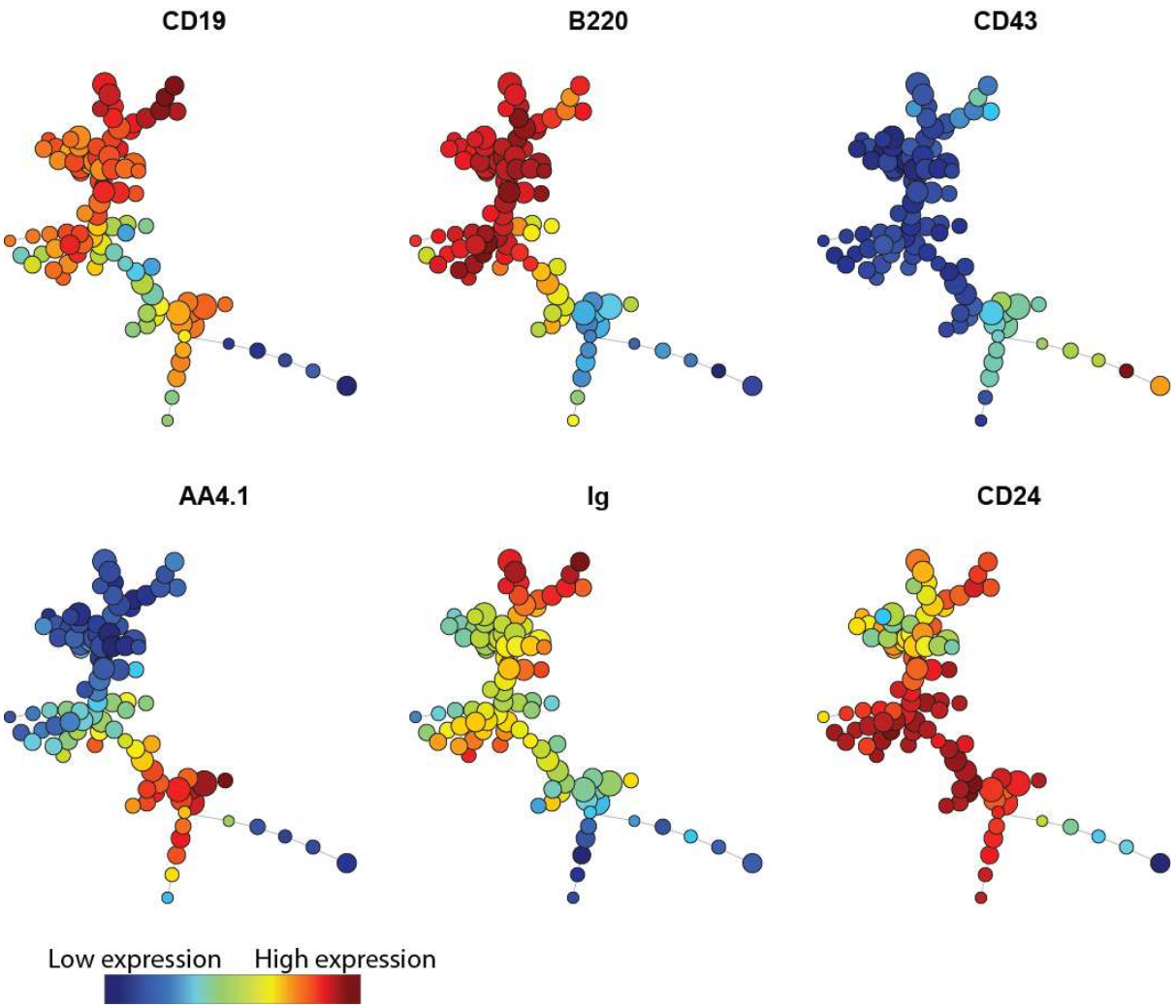


Figure 3-6: Marker expression of B cell fractions in CTG2 mice allows characterisation of intermediate populations.
CTG2 mouse FlowSOM minimal spanning tree from Figure 3-5, overlaid with the expression level of 6 cell surface markers. The scale of low and high expression is different for each marker. For example, the B220 marker is expressed in all cells, despite the range of low and high expression, because all cells used in the analysis were pre-gated for B220⁺ cells.

The number of cells in fraction A, D and intermediate EF are statistically different between WT and CTG2 (p-value 0.037, 0.037 and 0.037 respectively, Mann-Whitney/Wilcoxon rank sum test; Figure 3-8 a). The intermediate EF population and, especially, fraction BC have a higher coefficient of variation in CTG2 mice compared to WT, even though the difference in fraction BC is not statistically significant between the two mice. The TKO mice have been shown to have a developmental block at the stage where a pre-BCR is required for further progression (Figure 3-7) (fraction BC to fraction C'D) and would be also expected to have a block at the later BCR stage (fraction D to fraction E) if a recombined heavy chain rescue was performed. The transgene in CTG2 mice successfully restored developmental progression (Figure 3-7). The dependence of successful development on the pre-BCR and BCR highlights their importance. As a result, fraction BC and fraction E are the two stages where I would expect to see a difference in CTG2 mice due to the heavy chain only nature of their pre-BCR and BCR. The deletion of the C_H1 domain circumvents the retention of the heavy chain in the endoplasmic

reticulum via the BiP chaperones protein, allowing cells to progress (Zou et al., 2007). However, the deletion may also impair the ability of the surrogate light chain to bind to the heavy chain to form a functional pre-BCR, reducing developmental progression. The high variation in CTG2 fraction BC B cell numbers and their relatively higher proportions compared to WT supports this notion; however, the major impact can be seen downstream in fraction D. Ultimately, the additional knockout of the SLC genes would be desired to further our understanding of the role the SLC plays in heavy chain only pre-BCR development.

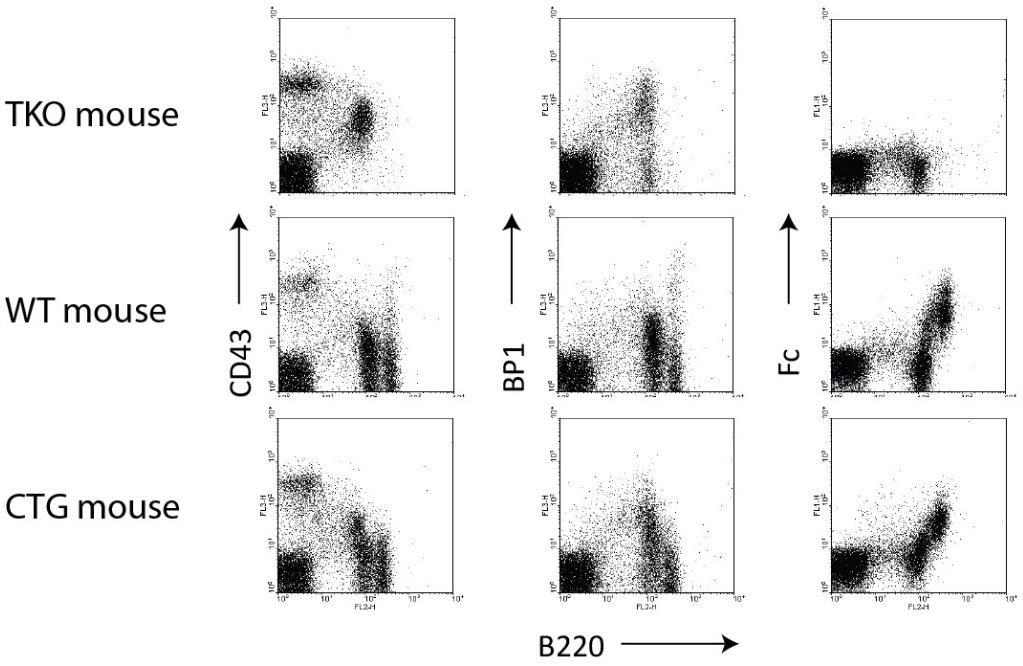


Figure 3-7: Flow analysis of TKO mice show B cell development does not move past fraction BC.
All B220⁺ cells in the TKO mouse are CD43⁺ and partly BP1⁺. No immunoglobulin (Fc) cell surface staining in TKO mouse shows B cells do not progress into the immature stage of development. Together this data points to a developmental block in the TKO mice at the fraction B/C - C', a stage that is reliant on the cell surface expression of the pre-BCR for developmental progression. The TKO block is successfully overcome in the CTG mouse and mirrors WT development. Figure adapted from data provided by Crescendo Biologics.

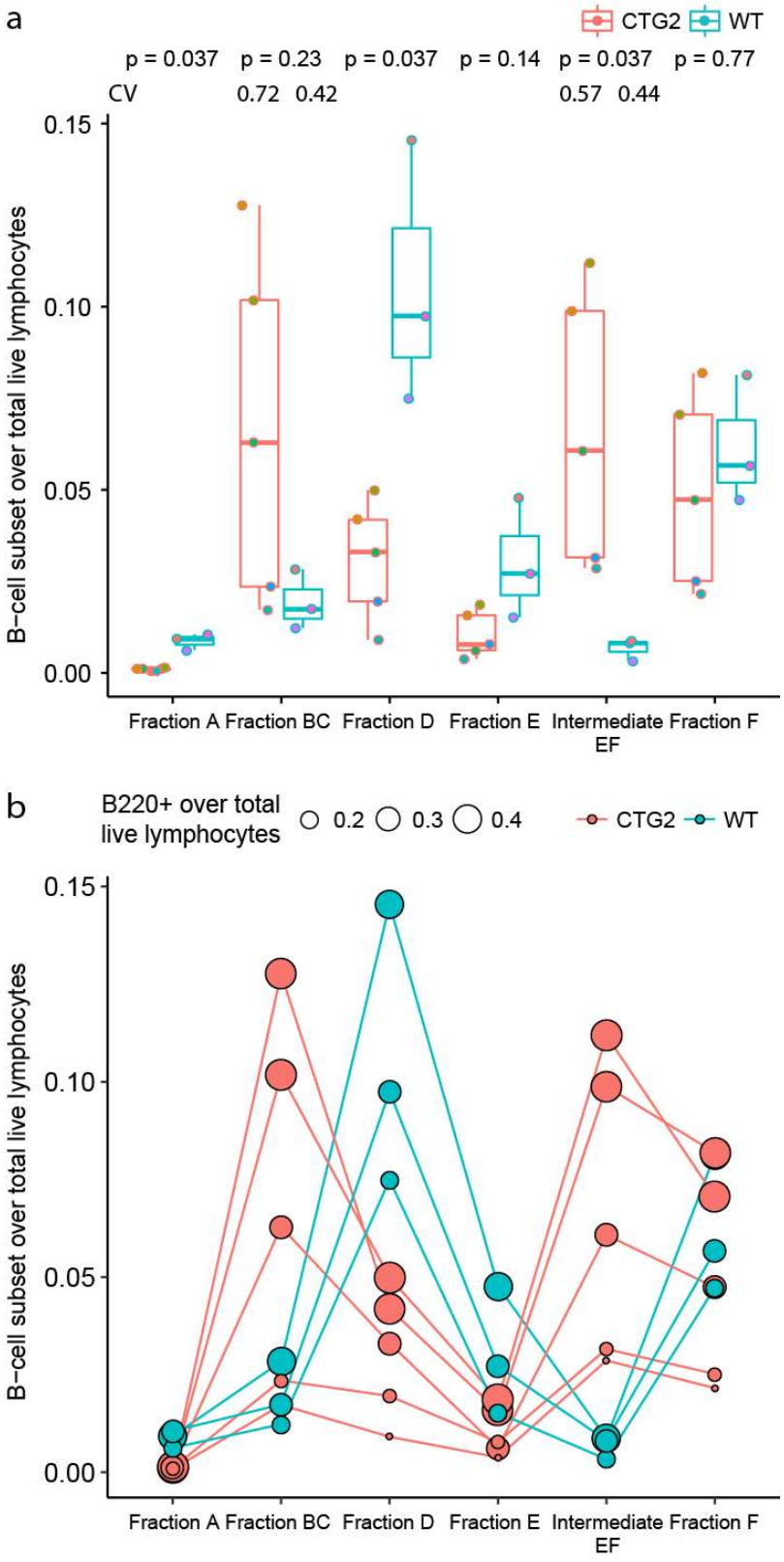


Figure 3-8: Quantification of the B cell fractions in WT (n=3) and CTG2 (n=5) mice.
(a) Proportions of B cell subsets over total live lymphocytes plotted for each fraction. Unpaired two-sample Wilcoxon rank sum test with continuity correction used to determine significance of difference between WT and CTG2 fractions. The coefficient of variation is shown for fraction BC and intermediate EF. (b) The line plots allow tracking of changes in B cell numbers between the developmental fraction for each sample, highlighting that the decrease in fraction D numbers is still present in samples with low initial fraction BC counts. Size of circles represents proportion of B220 positive cells over total live lymphocytes.

During the fraction D stage of development, large pre-B cells undergo approximately 6 rounds of cell division. This clonal expansion increases the pool of cells that subsequently undergo Igl recombination, increasing the overall repertoire diversity. The low number of fraction D cells suggests that this expansion does not take place, perhaps due to an impairment in pre-BCR signalling. Nevertheless, as the Igl is knocked out in CTG mice, the expansion is not required for repertoire diversification and by fraction F the B cell proportions are equivalent to WT levels (Figure 3-8 a). Tracking the fraction number in each sample reveals that high fraction BC levels in CTG2 mice also predicts high intermediate EF numbers (Figure 3-8 b). Interestingly, the B cell number in CTG2 mice still decrease as they transition from fraction D to E, as is the case in WT. This suggests that despite not undergoing Igl productive recombination and HC pairing selection, the CTG2 B cells still face survival pressures. Because the pre-BCR and the BCR in CTG mice should be the same heavy chain at both developmental fractions, the additional selection observed suggests that they require fine-tuned levels of tonic signalling to pass through both selection checkpoints.

Altogether, the analysis of the bone marrow developmental fraction reveals an expected decrease in fraction D B cells (Zou et al., 2007), while the overall and final fraction F numbers were in line with WT mice. With the use of unbiased analysis, an unexpected intermediate population between fraction E and F was discovered, which together with increased B cell numbers in fraction BC of CTG2 mice suggests that only certain selected heavy chain only pre-BCRs and BCRs go on to form the mature B cell pool.

3.3 Increased marginal zone B cells and reduced follicular B cells observed in CTG mouse spleen with reduced transitional 1 and 2 populations

I next wanted to examine whether the splenic B cell populations showed any deviations from the WT norm. The first noticeable difference with the CD19 B220 staining was a tail $B220^{low}$ population emerging from the B-2 B cells (Figure 3-9 a top panel). In WT, this population would be considered to consist of B-1a and B-1b cells. To determine if these cells correspond to B-1 B cells in CTG2 mice I quantified the $CD19^{+}B220^{-}CD23^{-}Ig^{+}$ cells. In CTG2, not all of the cells correspond to B-1 cells, suggesting that the lower B220 expression is a B-2 characteristic of B cells arriving from the bone marrow as noted above (Figure 3-4). Quantifying the proportion of B-1 cells over total lymphocytes did not reveal a significant difference between WT and CTG2 mice (Figure 3-9 b), although the dispersion of CTG2 B-1 counts was much greater than for WT (CV 0.58 CTG2 vs 0.18 WT). This suggests that some CTG2 mice do have an enlarged B-1 cell population. More apparent are the differences in the marginal zone (MZ) and the follicular B cell (FO) populations (Figure 3-9 a middle panel). CTG2 mice have a significantly higher MZ population and a lower FO B cell population (Figure 3-10 a). In addition, the transitional B cell populations T1 and T2 are significantly reduced in CTG2 mice (Figure 3-9 a bottom panel and Figure

3-10 b). The reduced levels of the T1-2 B cells suggests a reduced number of B cells are exiting the bone marrow. This notion could be supported by the large intermediate EF and fraction F Ig^- populations detected in the bone marrow (Figure 3-5 b), which point to a block in B cell development and a decreased migration into the spleen compartment.

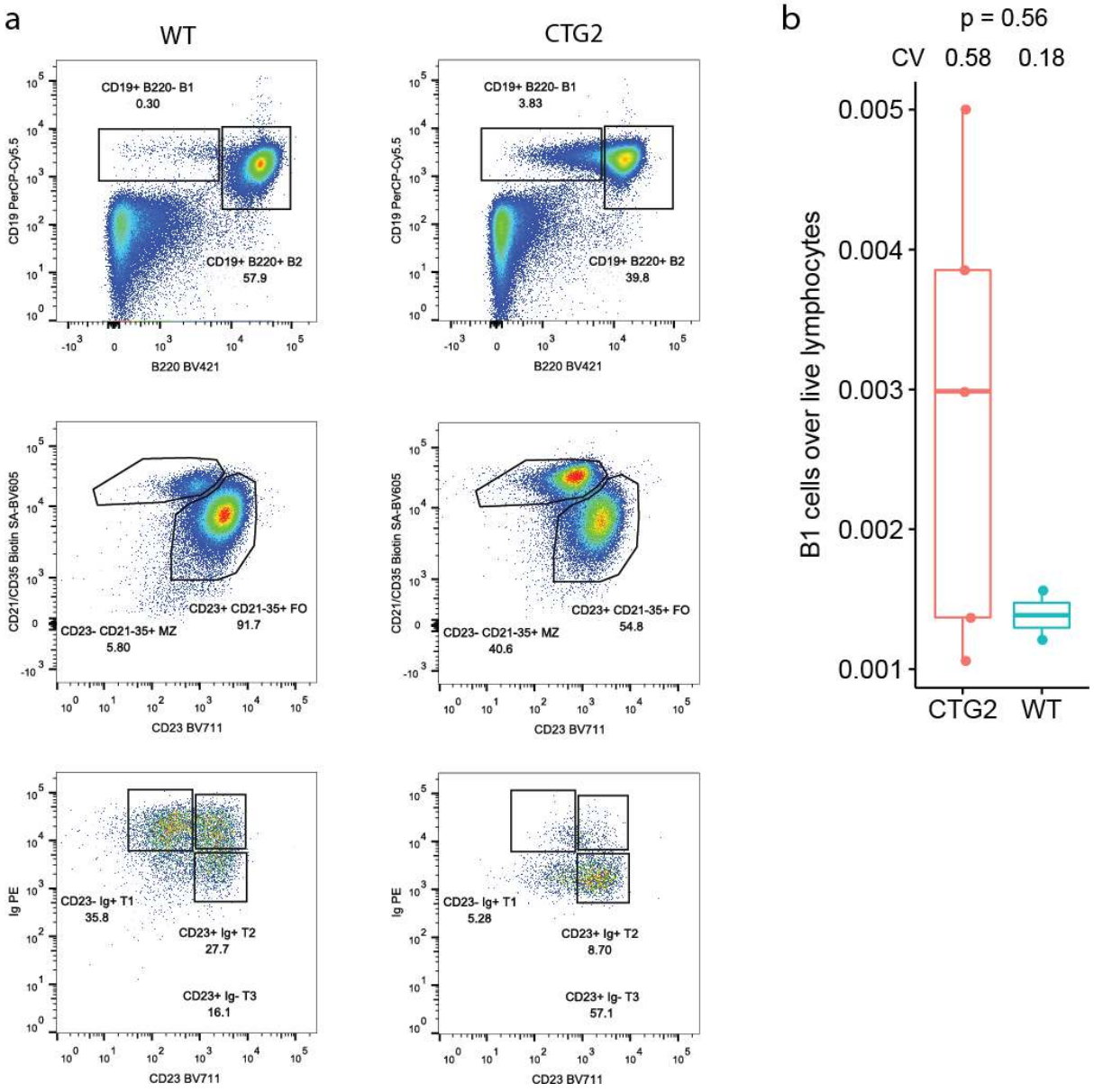


Figure 3-9: Analysis of WT and CTG2 spleen B cell populations.
(a) Top panel shows CD19 vs B220 demarcating B1 from B2 cells. The middle panel shows CD23 vs CD21/CD35 separating follicular (FO) B cells from marginal zone (MZ) B cells. Bottom panel shows CD23 vs Ig, highlighting the transitional 1-3 B cell populations. (b) Proportion of B1 cells over live lymphocytes in WT and CTG2 spleens. The coefficient of variation (CV) is provided above the plot, along with the p-value from the unpaired two-sample Mann-Whitney/Wilcoxon rank sum test with continuity correction.

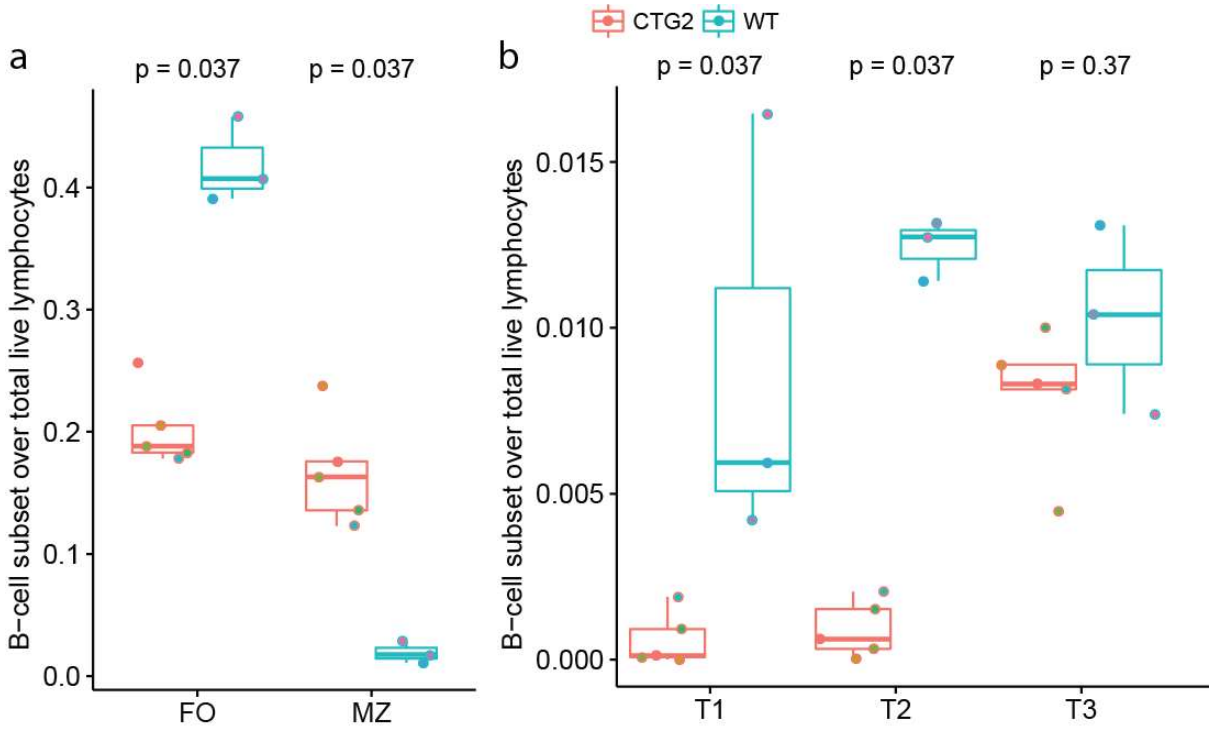


Figure 3-10: Quantification of spleen B cell population in WT (n=3) and CTG2 (n=5) mice.
P-values were calculated with an unpaired two-sample Wilcoxon rank sum test with continuity correction. (a) FO MZ populations. (b) T1-3 populations.

Altogether, these results show that the proportions of spleen B cell subsets are altered in CTG2 mice compared to WT, as expected from studies of $SLC^{-/-}$, $sIgM^{-/-}$ and models of IgG only isotype B cell development. Despite these alterations, the overall B cell compartment size still mirrors that of WT.

3.4

The following text has been redacted due to sensitivity of the material

3.4.1

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-11

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-12

The following text has been redacted due to sensitivity of the material

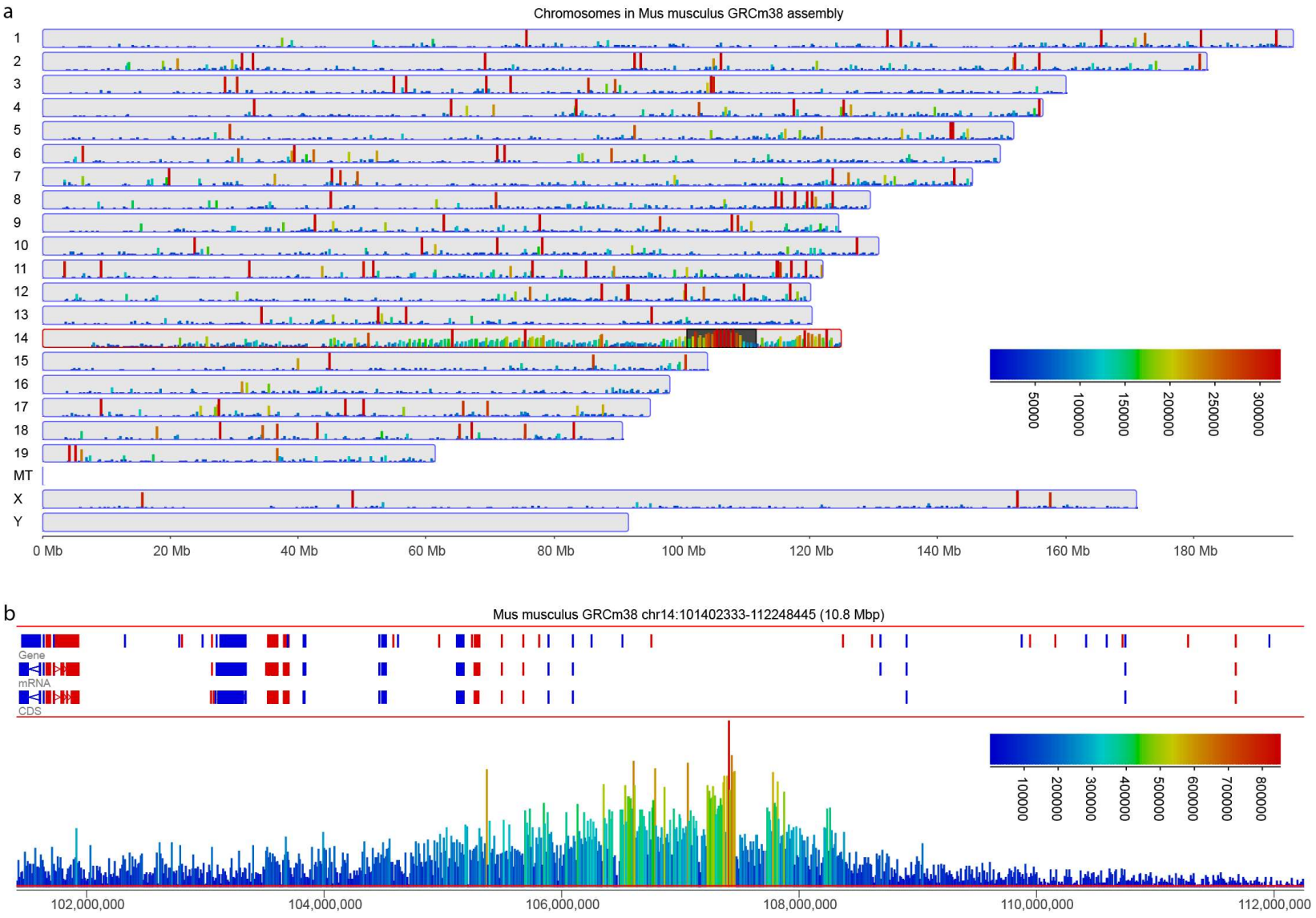
3.5 Targeted locus amplification reveals location of transgene incorporation and sequence composition

The Crescendo mouse was made by microinjection of the YAC transgene into mouse oocytes. This means the transgene integration into the genome is random and may disrupt endogenous genes and regulatory elements. To determine if any gene disruptions may have contributed to the phenotype observed by flow cytometry, I decided to determine the location of the transgene.

My initial attempts using techniques that relied on restriction enzyme digestion followed by PCR amplification of the transgene flanking regions were unsuccessful and prompted me to utilise a new technique reliant on proximity ligation, developed in Wouter de Laat's group, called targeted locus amplification (TLA) (Vree et al., 2014) (Figure 3-13).

Figure 3-13: Difference in read coverage obtained from TLA compared to 4C.
Adapted from (Vree et al., 2014).

TLA works on the principles of chromosome conformation capture (3C) techniques, namely 4C. The chromatin within each nucleus of a cell is digested using a restriction enzyme, followed by ligation. The result is the fragments that are close together in physical space, not linear genomic space, are ligated into DNA circles of co-localised restriction fragments. These large DNA circles are further fragmented with a secondary, less frequent cutter, and PCR amplified with a pair of outward facing anchor primers. This amplification enriches for the sequences of interest. Unlike in 4C, where only the DNA around the restriction cut sites is sequenced and analysed, in TLA the full restriction fragment is amplified and sequenced (Figure 3-13). This process reveals sequences adjacent to the anchor, much like the flanking region PCR amplification methods described above, but with greater distance and coverage. In order to be able to amplify the flanking regions of the transgene, I have previously used a pair of primers spanning the transgene YAC arms and the adjoining human/mouse sequence to verify the integrity of the transgene ends (data not shown). Because of the confirmed presence of this region of the transgene, I designed the TLA anchor primers within the Hyg HIS arm joint (Figure 3-2). This anchor region forms the 4C viewpoint from which the adjacent mouse sequence can be sequenced.



(legend on next page)

Figure 3-14: TLA reads aligned to the GRCm38 genome.
(a) Genome view of deduplicated TLA reads. Chromosome 14 contains a broad distribution of reads centring on the highlighted region (black box). (b) Chromosome view of highlighted region showing the possible location of the transgene insertion. The highest peak is over a gene poor region surrounded by snoRNA genes.

Much like in 4C, it is expected that the increase in genomic distance from a viewpoint results in a decay of contact probability, or in the case of TLA a decay in read coverage. Using this principle, we can use TLA reads mapped to the mouse genome to ascertain a possible location of the transgene insertion. As part of the analysis, a set of blacklisted regions produced by the ENCODE project corresponding to regions with artificially high levels of signal were empirically removed from further analysis (The ENCODE Project Consortium, 2012). The relatively high coverage observed on chromosome 14 suggests that the transgene is located here (Figure 3-14 a - black box highlight). A zoom in on this region reveals a peak of reads centring within a gene poor region and surrounded by snoRNA genes (Figure 3-14 b).

Thus, the TLA revealed that the transgene integration into a gene poor site prevented the disruption of any functionally relevant genes. In addition, the functional recombination and antibody production observed in CTG2 mice means the heterochromatin that is usually associated with gene poor regions, such as the proposed integration site, did not spread and impact the activity of the transgene (Wang et al., 2014).

3.5.1 3D DNA FISH with chromosome paints confirms TLA location of transgene

3.5.1.1 Human V-D intergenic region DNA FISH probe design

The likely inability of the HCAbs to pair with the SLC may impose additional constraints on the productive recombination events that will lead to signalling and developmental progression. To test this, I investigated the allele recombination frequency in CTG2 B cells to determine if the frequency of productive first round recombination was similar to WT, or if the second allele was required to recombine at a higher frequency. Alteration in the expected number of double allele recombiners could provide a possible explanation for the higher number of fraction BC B cells observed in CTG2 mice, as these cells would need to spend more time at this developmental stage in order allow their second allele to recombine. To accomplish this, I needed to design and manufacture probes for the human VD intergenic region, similar to the mouse VD1-7 probes described previously (Ramadani et al., 2010). The human VD intergenic region is considerably smaller than mouse and in order to maximise the size of the probes, I chose to use the DNA FISH probe design tool webFISH (Nedbal et al., 2012). WebFISH also considers repeats that are unique to the designated region for probe design, giving it an advantage over a manual approach with RepeatMasker. WebFISH returned four probes for the VD intergenic region ranging from 5kb to 500bp (Figure 3-15).

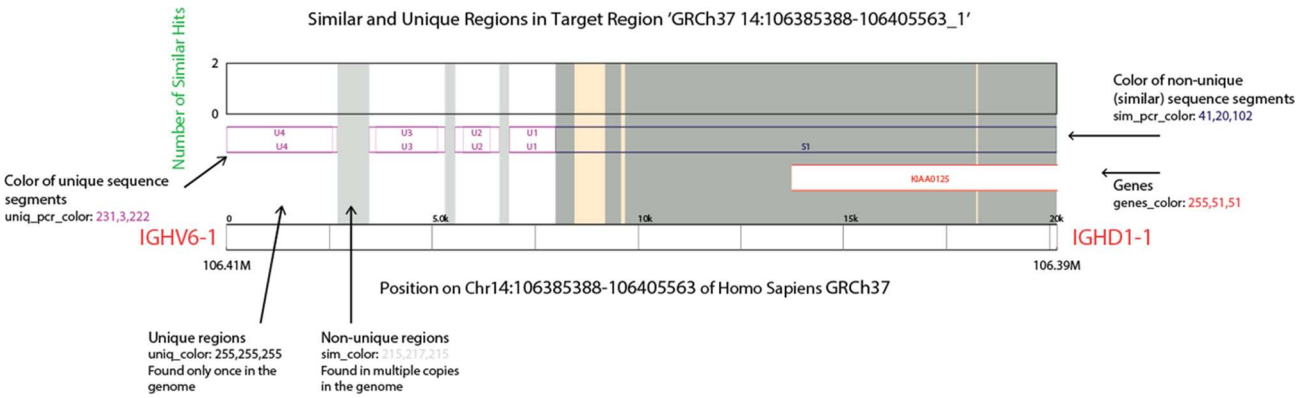


Figure 3-15: Graphical output from webFISH showing the four designed probes.
Four probes were generated for the human V-D intergenic region, shown as purple boxes. The non-unique regions are highlighted in grey. Red box shows the KIAA gene while the blue box highlights the non-unique similar region.

3.5.1.2

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-16

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-17

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

3.5.1.3

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-18

The following text has been redacted due to sensitivity of the material

3.5.2 Transgene sequencing allows correction of the reference sequence

The transgene reference sequence, to a large extent, is based on the mouse reference genome sequence onto which the end-sequences of BACs, used in the construction of the transgene YAC, were mapped. Because the human genome reference was built from a mixture of individuals and the human sequences within the BAC will naturally contain polymorphisms, it is expected that the transgene reference will contain single nucleotide polymorphisms (SNPs).

By aligning TLA reads to the transgene reference, single nucleotide mismatches composed of a single letter consensus can be observed throughout the transgene. In some instances, the TLA reads also highlight several nucleotides missing from the reference (Figure 3-19). More interestingly, there are also regions of low coverage, which can be filled using local alignment (Figure 3-20). Taking such mismatches into consideration for future work can prove invaluable with, for example, CRISPR-Cas9 editing of the integrated transgene, since even a single mismatch in the guide RNA targeted sequence can have a negative impact on editing efficiency (Zheng et al., 2017).

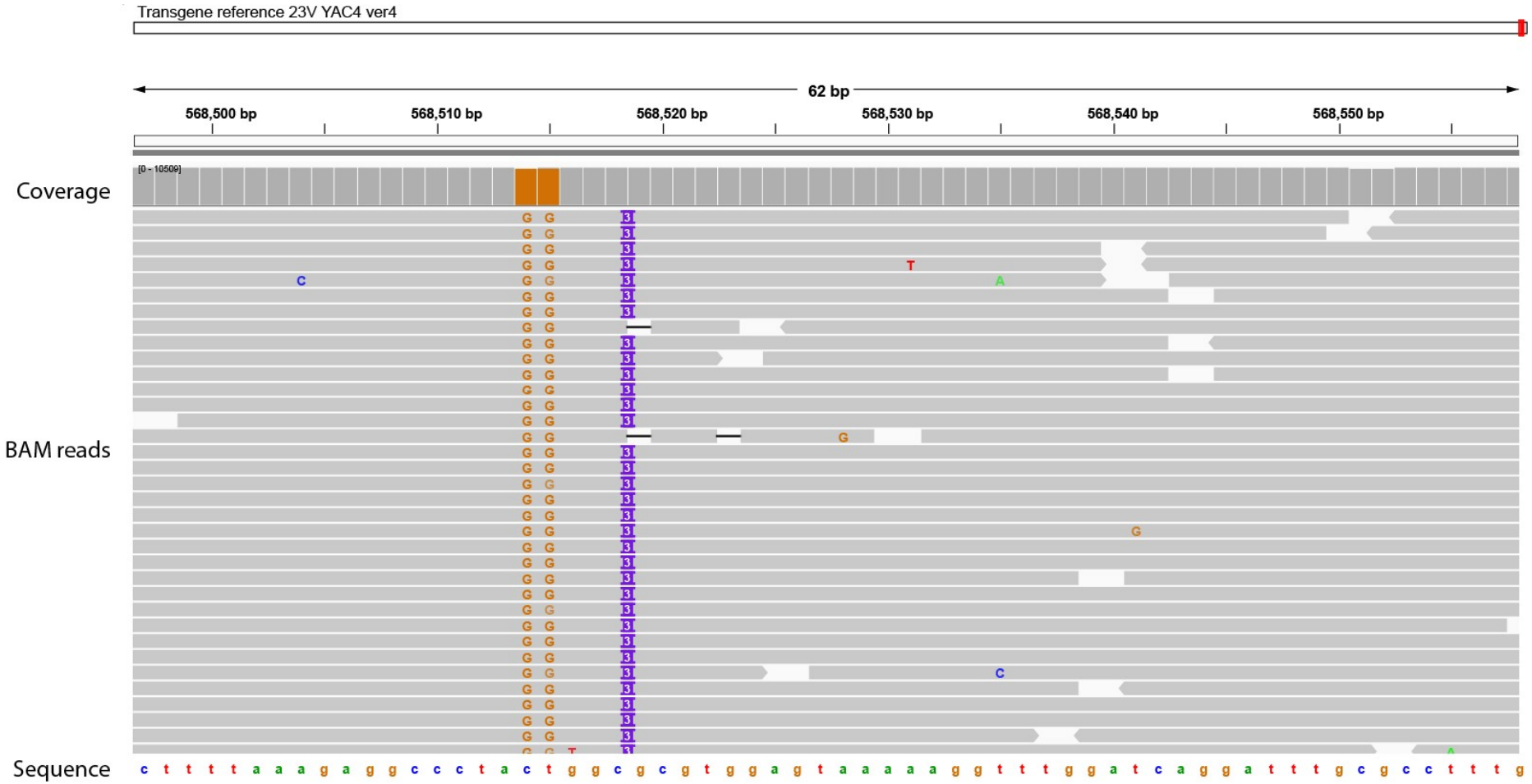


Figure 3-19: Expected SNPs in TLA data.
An example of single nucleotide polymorphisms along with a three basepair (insertion in purple) absent from the transgene reference.

(legend on next page)

Figure 3-20: TLA reveals regions differing from the reference sequence.
End-to-end alignment of TLA reads to the transgene reference sequence produces regions of low coverage. Using local alignment, the TLA reads are trimmed/masked which allows parts of the reads (local) to be aligned. The bases that were not aligned to the reference are shown over the gap. The consensus formed by these locally aligned reads provides a way of correcting these small gaps within the reference.

3.5.3

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-21

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-22

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-23

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-24

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-25

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-26

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-27

The following text has been redacted due to sensitivity of the material

3.6 CTG2 mice primarily utilise distal V genes and have a reduced proportion of productive recombination in fraction BC that recovers in fraction C'DE

I have previously quantified recombination in CTG2 B cells merging fraction BC and C'DE together (Figure 3-25). I now wanted to examine the recombination of the two developmental stages independently. The random incorporation of nucleotides during the joining of the VDJ gene segments can result in out-of-frame rearrangements with a theoretical 2 in 3 chance due to the triple nature of codons. From previous experimental data, we know that the proportion of productive to unproductive rearrangements in mouse fraction BC is around 1/3 and goes up to 2/3 in fraction D (Bolland et al., 2016; Ehlich et al., 1994). The same proportions are observed in our mouse WT VDJ-seq data. Interestingly, the proportions of productive to unproductive of the human IGH transgene in CTG2 mouse at fraction BC are considerably lower compared to mouse WT Igh (11% CTG2, 31% WT) (Figure 3-28). By fraction D, the proportion of productive sequences recovers, reaching almost WT levels (54% CTG2, 65% WT).

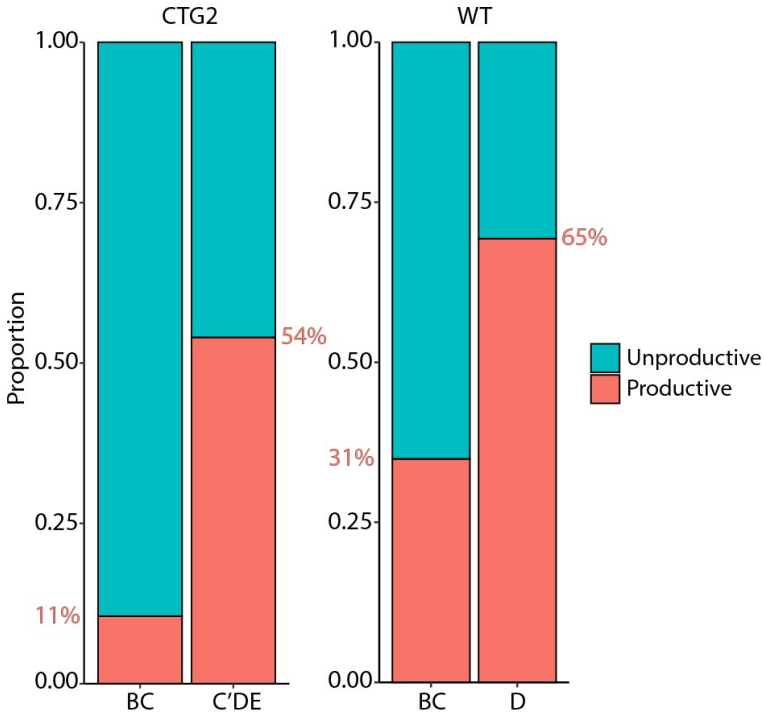


Figure 3-28: Productive recombination in CTG2 fraction BC is lower than expected, but recovers in fraction C'DE.
CTG2 mouse with human transgene recombination on the left (mean of BC n=3, C'DE n=2) and WT mouse on the right (mean of BC n=3, D n=3). Daniel Bolland generated WT VDJ-seq data.

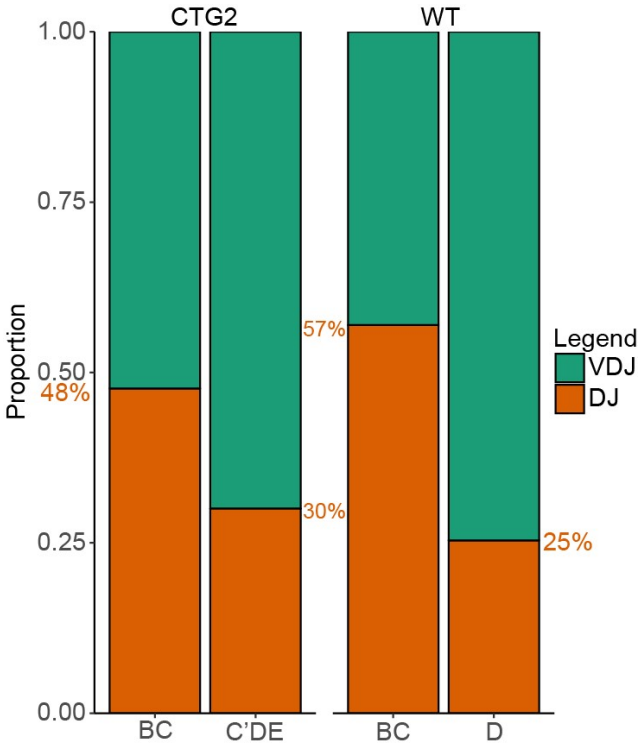


Figure 3-29: Proportion of DJ:VDJ recombination is lower than expected in CTG2 fraction BC, but is re-established in fraction C'DE.
Mean proportions of CTG2 mice are plotted on the left (BC n=3, C'DE n=2) with WT on the right (BC n=2, D n=2). The ratio is consistent between biological replicates (CTG2 BC VDJ - 0.55, 0.50, 0.52; BC DJ - 0.45, 0.50, 0.48; CDE VDJ – 0.70, 0.70; CDE DJ – 0.30, 0.30; WT BC VDJ – 0.44, 0.42; BC DJ – 0.56, 0.58; D VDJ – 0.76, 0.73; D DJ – 0.24, 0.27) . Sam Rees generated the WT datasets.

The high proportion of unproductive recombination events at fraction BC in CTG2 mice suggests that a large proportion of B cells will be required to recombine their second allele. To test this hypothesis I examined the proportion of VDJ to DJ recombination events in both fraction BC and C'DE. The proportion of VDJ to DJ recombination events in fraction BC has been previously shown to be around 66% DJ to 34% VDJ (Bolland et al., 2016). If more B cells need to recombine their second allele, the proportion of detected DJ recombination events should be lower. In WT fraction BC, only 57% of reads were identified as DJ, suggesting that perhaps some of these cells may be from a later fraction. Nonetheless, CTG2 fraction BC cells displayed a lower than expected proportion of DJ reads at 48% (Figure 3-29). A prerequisite for B cells that progress into fraction D is a productive recombination that can be translated into a protein pre-BCR. Therefore, it can be expected that all B cells in fraction D will have one productively recombined VDJ allele with a possible combination of a DJ or an unproductive VDJ on the other allele. This leads to the ratio of 75% VDJ to 25% DJ that I observed in WT fraction D (Figure 3-29) and that also very closely matches the first reported proportion in transformed mature B cells, where 40% of cells were shown to have recombined both alleles (70:30 VDJ:DJ) (Alt et al., 1984). The proportion in CTG2 fraction C'DE is very close to the WT, confirming the ability of the HCAb

receptors to support progression from the pro-B (fraction BC) to the pre-B (fraction C'D) stage of B cell development, despite generating fewer productive recombination events in pro-B cells.

Table 3-1: Proportion of VDJ and DJ calls in VDJ-seq data.
Proportions were rounded to two decimal places.

VDJ calls	DJ calls	Fraction	Sample type	VDJ:DJ	
0.67	0.21	D	WT	0.76	0.24
0.69	0.25	D	WT	0.73	0.27
0.38	0.49	BC	WT	0.44	0.56
0.36	0.49	BC	WT	0.42	0.58
0.16	0.13	BC	CTG2	0.55	0.45
0.20	0.20	BC	CTG2	0.50	0.50
0.18	0.16	BC	CTG2	0.52	0.48
0.25	0.11	CDE	CTG2	0.70	0.30
0.26	0.11	CDE	CTG2	0.70	0.30

I additionally wanted to ascertain that recombination between the transgene and the endogenous mouse Igh, which still had all VDJ genes, was not taking place. I first checked if merging the human and mouse IMGT reference would allow me to distinguish between the two species. The Levenshtein distance between human and mouse VDJ genes was great enough that it should be possible to get high quality calls for species chimeric recombination (Figure 3-30). I did not find any such reads, confirming that recombination is limited to within the CTG2 transgene.

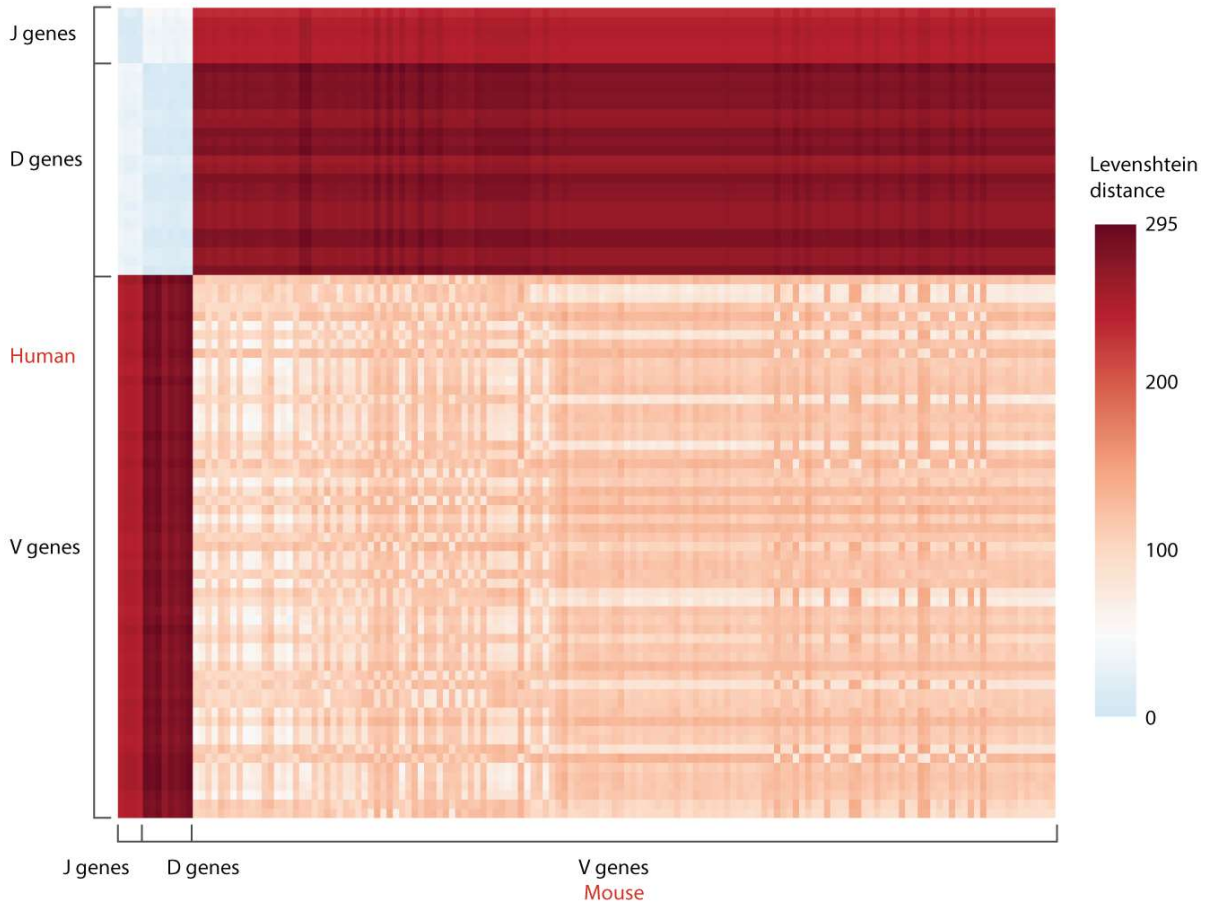


Figure 3-30: Heatmap of the Levenshtein distances between human and mouse VDJ genes of the IGH.

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-31

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following figure has been redacted due to sensitivity of the material

Figure 3-32

The following text has been redacted due to sensitivity of the material

3.6.1 CTG2 mice do not show signs of autoreactive antibodies

Previous studies suggest that autoreactive BCRs tend to have longer CDR_H3 regions compared to non-autoreactive BCRs and that B cells bearing these BCRs tend to be selected against in humans, favouring shorter CDR_H3 in the periphery (Kaplinsky et al., 2014; Larimore et al., 2012; Wardemann et al., 2003). To further explore the possibility of autoreactive receptors being present within the HCAb repertoire, I examined the length and amino acid (aa) composition of the CDR_H3. The CDR_H3 median length in CTG2 mice matches previously reported human productive CDR_H3 lengths (14 aa) (Zemlin et al., 2003). Surprisingly, the more recent repertoire sequencing study of large pre-B cells from human bone marrow (shown above in Figure 3-32) showed the average length of CDR_H3 to be 2 aa longer at 16 aa (Martin et al., 2016) (Figure 3-33) (the difference is present despite the exclusion of the C and W conserved residues from both datasets). The length distributions of productive and unproductive recombination were almost identical in CTG2 mice and the mean CDR_H3 size of productive recombination, on which selection would act, was only marginally higher in fraction C'DE compared to fraction BC (Table 3-2). In mouse early stages of development a selection towards longer CDR_H3 was observed in the V_H7183 family (Ivanov et al., 2005), which supports the preferential selection of marginally longer CDR_H3 in CTG2 B cell fraction C'DE (Table 3-2). In contrast, analysis of human repertoires, with the use of unproductive recombination as a proxy for pro-B cells (fraction BC), showed a preferential selection of shorter CDR_H3 sequences (Larimore et al., 2012). This suggests that the preferential selection of longer CDR_H3 is a result of the mouse context rather than a feature of the human sequence.

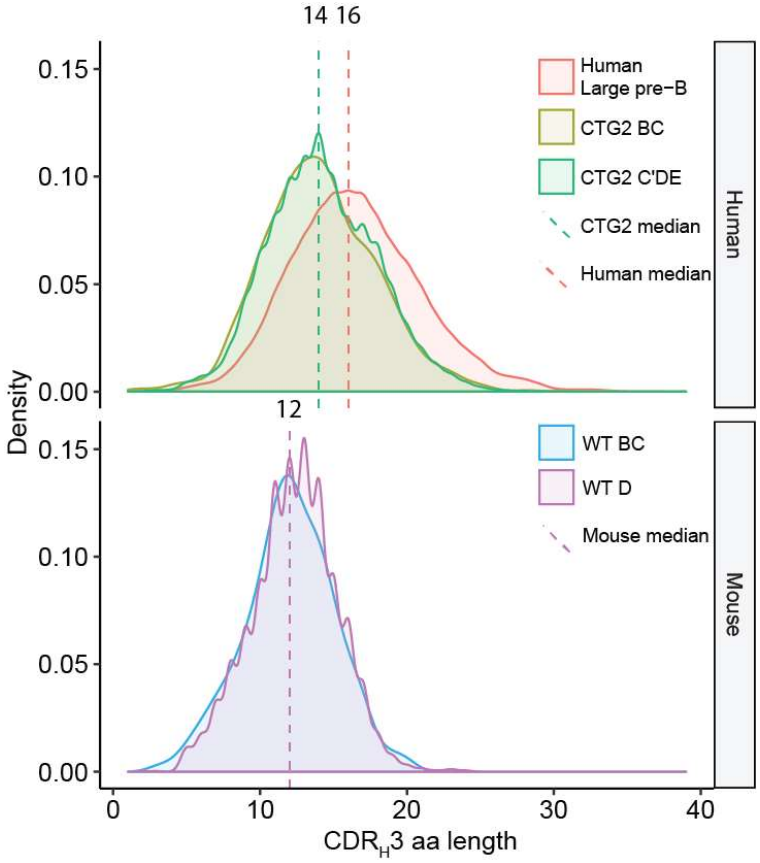


Figure 3-33: CDR$_H$3 size distribution of productive recombination events.
Dashed lines indicate the median of the distributions. Human large pre-B cell data were obtained from PRJNA39946 study (Martin et al., 2016).

Table 3-2: Mean length of CDR_H3 in productive and unproductive recombination sequences.

Productive		**Unproductive**	
Cell type	Mean CDR_H3 length	Cell type	Mean CDR_H3 length
Human large pre-B	16.54	Human large pre-B	16.43
CTG2 BC	14.09	CTG BC	14.01
CTG2 C'DE	14.26	CTG C'DE	13.95
Mouse WT BC	12.03	Mouse WT BC	12.55
Mouse WT D	12.32	Mouse WT D	12.41

Previous studies have shown that autoreactive antibodies have a preferential usage of positively charged amino acids, namely the frequent appearance of arginine (Barbas et al., 1995; Köhler et al., 2008). I next checked whether HCAbs in CTG2 show preferential amino acids usage. As noted in previous comparisons of human and mouse antibody sequences, tyrosine (Y) is preferentially utilised in mice and somewhat more frequently in CTG2 mice compared to human (Figure 3-34 left and middle) (Zemlin et al., 2003). More importantly, arginine (R) is less prevalent in CTG2 then in WT, while histidine (H) and lysine (K) are more frequent in CTG2 CDR_H3. These residues are only marginally different from the human large pre-B CDR_H3, suggesting that the differences observed are due to species sequence differences rather than selection (Figure 3-34 right). Other residues such as cysteine (C), isoleucine (I), threonine (T), valine (V), proline (P), methionine (M) and glutamic acid (E) that have been shown to be significantly higher in humans compared to mice are also elevated in CTG2 compared to WT (Zemlin et al., 2003).

Altogether, these results suggest that the HCAb repertoire is not preferentially selected for autoreactive antibodies. The deleted C_H1 domain in CTG2 BCRs that alters the proliferative boost at the pre-BCR developmental stage, seems to still enable sufficient signal transduction to sustain differentiation into mature B cells without the reported self-specificity of the WT pre-BCR (Köhler et al., 2008; Meixlsperger et al., 2007) (section 1.2.2.3).

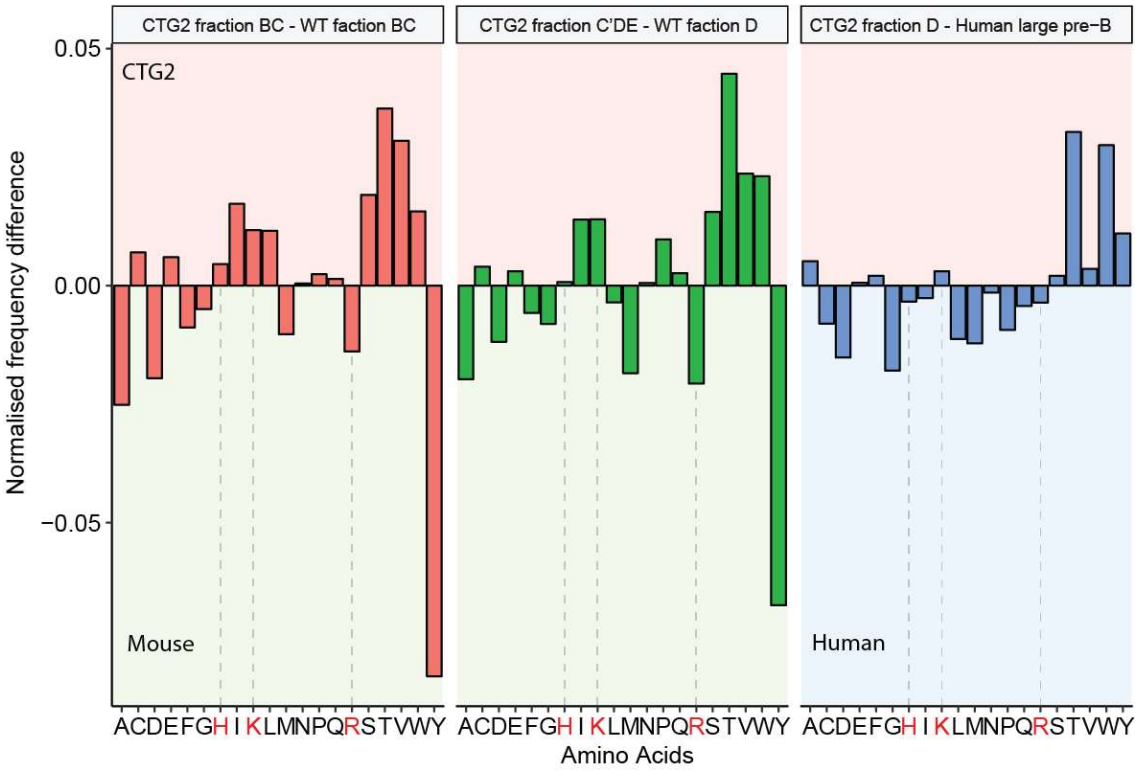


Figure 3-34: Comparison of amino acid usage between CTG2 mice and mouse/Human.
Positively charged amino acids (aa) are highlighted in red and a dashed line. Frequency of each aa was normalised to the total aa count for each sample. Human large pre-B cell data were obtained from PRJNA39946 study (Martin et al., 2016).

3.6.2 The CTCF binding site proximal to V gene RSSs are found at recombination active genes in both CTG2 mice and humans

We have previously shown that the recombination information content (RIC) scores that describe the potential of an RSS to recombine, act more like a binary switch, while other chromatin factors influence the recombination frequency of individual V genes (Bolland et al., 2016). To examine whether RSS RIC scores can explain the recombination frequency of individual human IGH V genes, I compared the V gene recombination frequency of large pre-B cells against their respective RIC scores. Because pre-B stage B cells have already undergone selection, I have only used unproductive events whose frequency should not have been impacted by selection, as they are only passenger events at this point (Kaplinsky et al., 2014; Larimore et al., 2012). The analysis reveals that much like in mouse, the recombination frequency is not fully explained by the individual V gene RIC scores. Most V genes have a high RIC score above -30, meaning their RSS sequence is close to the consensus, and there is no clear correlation between RIC score and V gene recombination frequency, either at the highly recombining or poorly recombining end of the spectrum. This prompted me to examine the chromatin status proximal to human RSSs.

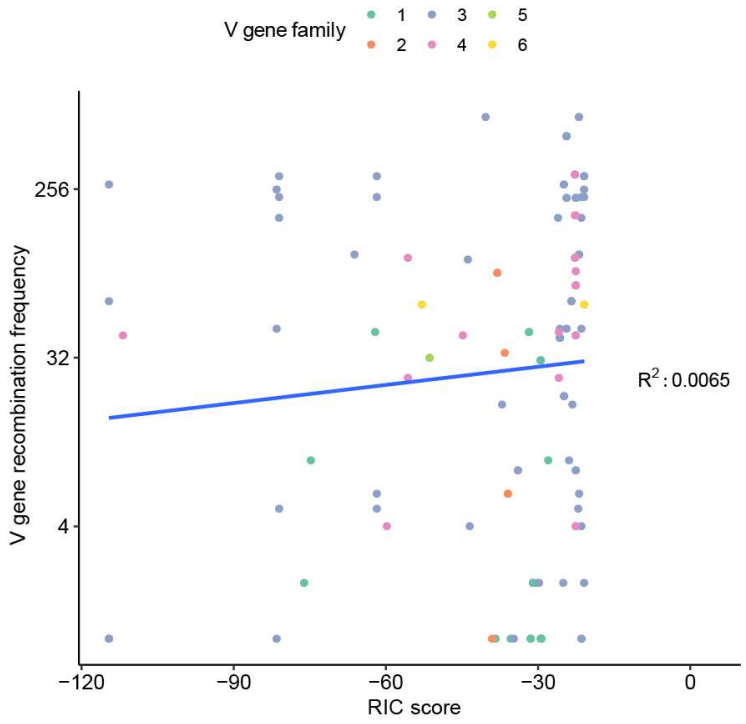


Figure 3-35: The recombination rate of individual V genes is not fully explained by RSS RIC scores.
RIC score is of the 23-spacer RSSs in humans. RSS scores were obtained from the Recombination Signal Sequence Site (https://www.itb.cnr.it/rss/). The recombination frequency of individual V genes were obtained from human bone marrow large pre-B cells (Martin et al., 2016).

There are no available ChIP-seq datasets for human pro-B (fraction BC) bone marrow cells. However, CTCF tends to have a highly conserved set of binding sites between different tissues (Holwerda and de Laat, 2013). Because of this conservation, the available CTCF binding sites from a human lymphoblastoid cell line are likely to reflect the sites in pro-B cells and can act as a proxy dataset. As a result, I focused my analysis on CTCF. Much like for the mouse Igh locus, CTCF binding proximal to human IGH V gene RSSs is enriched in evolutionary conserved clans II and III (Figure 3-36 top). This suggested that the chromatin states uncovered for the mouse Igh may be conserved across species and also influence recombination in humans (Bolland et al., 2016). The overlapping binding of cohesion (RAD21) with CTCF (Holwerda and de Laat, 2013) also prompted me to analyse the RAD21 ChIP-seq dataset and examine if the same pattern of clan segregation observed with CTCF held true for RAD21. Indeed, clan II and clan III V gene RSSs displayed a more proximal binding pattern compared to clan I genes, although to a lesser degree than CTCF (Figure 3-37). Interestingly, the highest density of functional V genes is at the same distance as clan II and III V genes (approximately 2^8-1; 255 bp), suggesting that there might be a relationship between CTCF and recombining V genes (Figure 3-36).

To examine the relationship of CTCF distance and functional V genes, I first plotted the frequency of unproductive recombination in large pre-B cells against CTCF distance. The clan II and clan III genes that mostly centred around 2^8 from the RSS turned out to be the dominant recombining V genes

(Figure 3-38 a), suggesting CTCF proximity is an evolutionary conserved feature of these V genes. The V genes used in the human CTCF analysis were limited to the V genes present within the CTG2 transgene. This was to allow a comparison with the human transgene within CTG2 mice, which provides a unique opportunity to extend the relationship between CTCF and V gene recombination frequency to other more transient chromatin factors such as PAX5 and EBF1. To ensure that CTG2 mice are a good model for other chromatin marks, I first wanted to validate that the RSS CTCF proximity within CTG2 mice still maintains the relationship with recombining V genes observed in humans, despite the differences in recombination frequency observed between the transgene and human pre-B cell unproductive events (Figure 3-32). Indeed, clan II and clan III V genes within the transgene that were recombining at a frequency above expected displayed the same relationship with proximal CTCF sites as the human recombination frequency data (Figure 3-38 b).

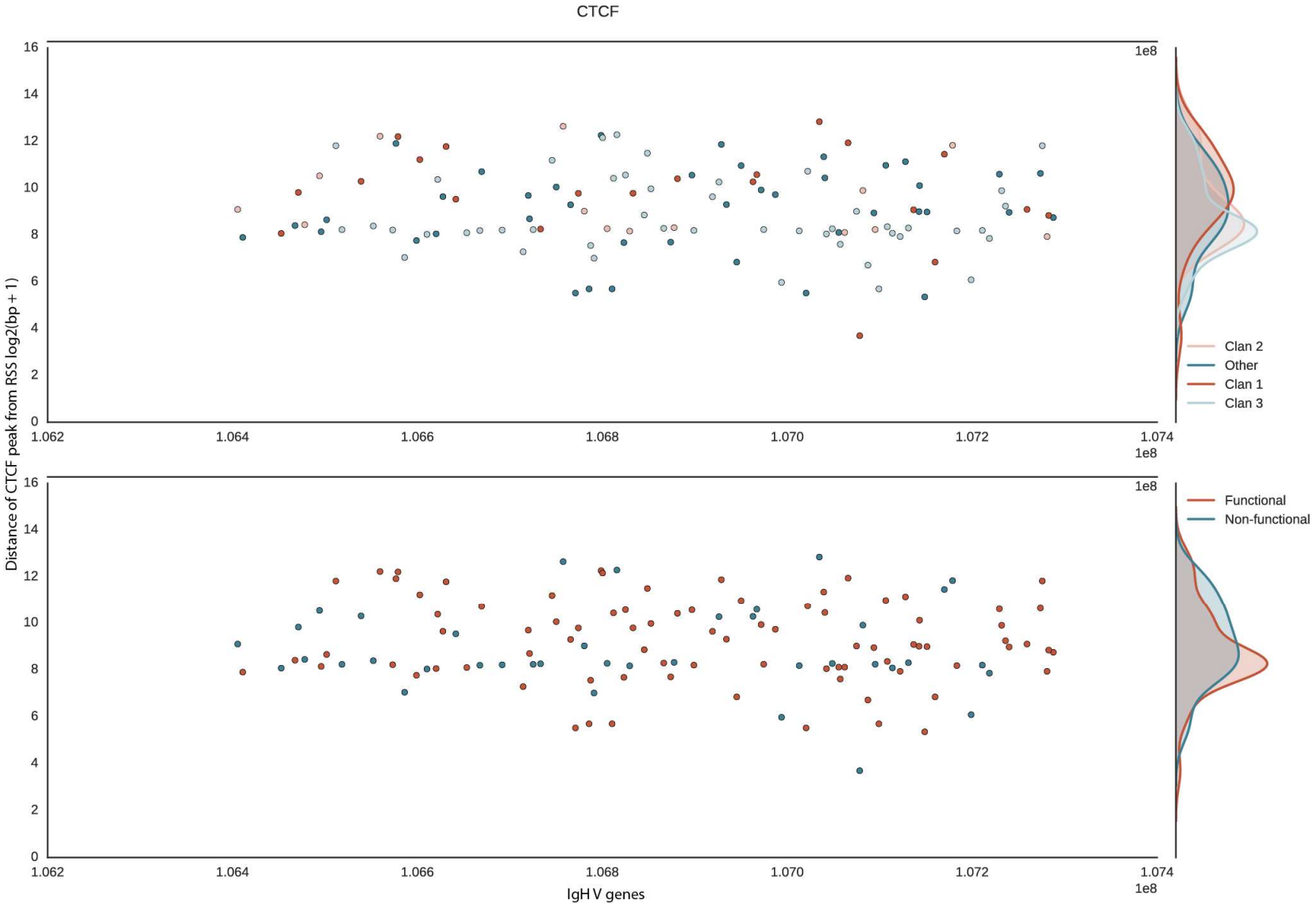
The following text has been redacted due to sensitivity of the material

Altogether, the pattern of CTCF suggests that the chromatin signatures matching active recombination within the human IGH are also present within the transgene. This makes it likely that the same Igh chromatin signatures uncovered within mouse pro-B cells are also implicated in the recombination of the human IGH.

Table 3-3: CTG2 mouse phenotype summary table.

Bone marrow B cell development	**Spleen B cell development**	**V(D)J recombination**
↑ variation in fraction BC B cell counts	↓ transitional B cells	↓ proportion of productive to unproductive fraction BC
↓ fraction D B cells	↓ follicular B cells	◯ fraction C'DE productive to unproductive
↑ intermediate fraction EF B cells	↑ marginal zone B cells	↓ proportion of DJ to VDJ recombination fraction BC
◯ fraction F B cell counts	↑ granularity and IgG cytoplasmic and vesicle staining	◯ fraction C'DE DJ to VDJ
	◯ B220/live B cells	
	◯ weight of spleens	

↑ Increased ↓ Decreased ◯ Same as WT



(Legend on next page)

Figure 3-36: Clan II and Clan III human V genes RSSs have proximal CTCF ChIP-seq peaks compared to Clan I.
(Top) Each V gene is assigned to one of three clans. If a V gene is not a member of any of three clans it is assigned to the 'Other' category. Y-axis marginal density plots highlight the distribution of distances for each clan. (Bottom) Each V gene is categorised as functional or non-functional based on the presence of stop-codons or defects in the regulatory elements of the germline sequence. More details about the generation of the figure are within the methods section.

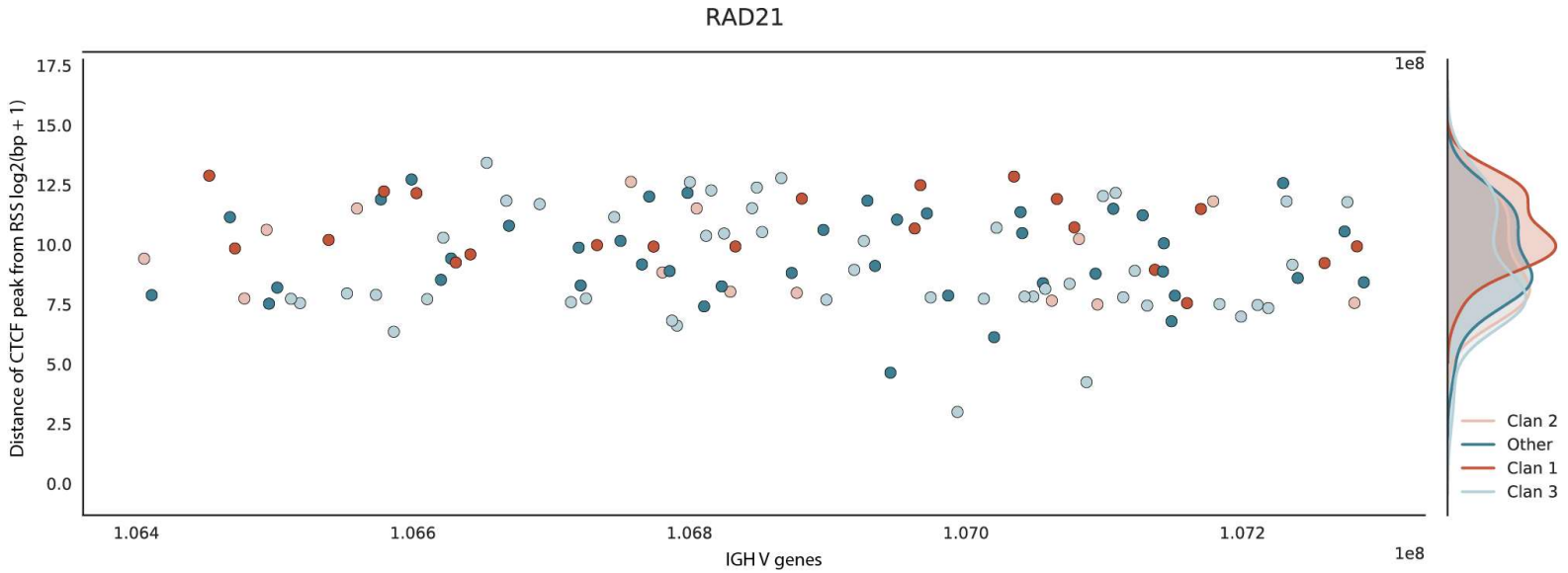


Figure 3-37: RAD21 ChIP-seq peaks display the same pattern of proximity to clan II and clan III V gene RSSs as CTCF.
Each V gene is assigned to one of three clans. If a V gene is not a member of any of three clans it is assigned to the 'Other' category. Y-axis marginal density plots highlight the distribution of distances for each clan. RAD21 data are from the GM12878 cell line.

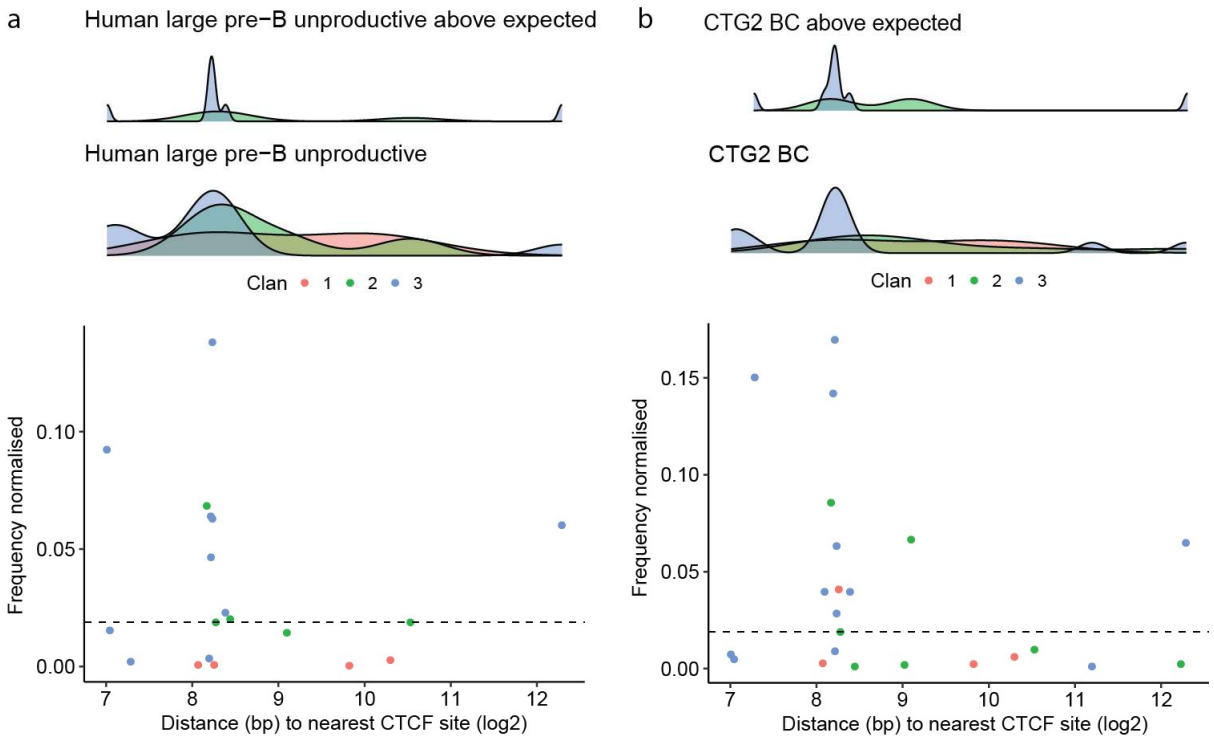


Figure 3-38: Distance of CTCF peaks from V gene RSSs is conserved for recombining V genes.
Horizontal dashed line indicated the expected frequency based on an even distribution of reads between recombining V genes. Top density plot shows distribution of CTCF distances for V genes above the expected threshold. Bottom density plot shows the CTCF distance distribution for all genes. CTCF data are from the GM12878 cell line. (a) Unproductive recombination frequency in human large pre-B cells. (b) CTG2 recombination frequency as the sum of both productive and unproductive events.

The following figure has been redacted due to sensitivity of the material

Figure 3-39

The following text has been redacted due to sensitivity of the material

3.7 Discussion

The use of dimensionality reduction techniques allowed the unbiased analysis of flow cytometry data and characterisation of B cell development in an HCAb human transgene model. Accumulation of B cells was observed at two points during bone marrow B cell development. Fraction BC showed a higher number of B cells, while fraction D was severely diminished compared to WT. The expression of the pre-BCR at fraction BC represents a major developmental checkpoint. Previous studies with $SLC^{-/-}$ mice and pre-BCRs restricted to the cytoplasm showed that these alterations resulted in a leaky pre-BCR phenotype and allowed developmental progression with diminished B cell numbers (Conley and Burrows, 2010; Ren et al., 2015; Shimizu et al., 2002). A possible cause of the developmental block and accumulation of B cells was revealed by repertoire analysis of fraction BC B cells. The repertoire analysis revealed a larger proportion of unproductive recombination than expected at this stage of development. This suggests more cells had to undergo recombination of the second allele. Indeed, the lower proportion of DJ reads in fraction BC supports this notion. Previous studies have shown that B cells with both unproductively recombined alleles accumulate in fraction C (Ehlich et al., 1994). The accumulation of B cells in CTG2 mice at fraction BC could be explained by a higher number of cells failing to recombine either of the two alleles.

Curiously, the cause of the increased unproductive recombination is not known and had not been previously documented in any autoreactive, $SLC^{-/-}$, $LC^{-/-}$ or HC only models (von Boehmer and Melchers, 2010; Ren et al., 2015; Shimizu et al., 2002; Zou et al., 2005, 2007). Interestingly, it has been suggested that the recombination machinery of mice works less efficiently on human VDJ genes, based on xenomouse models (mice with fully human antibodies) (Longo et al., 2017). Studies of $SLC^{-/-}$ mice also revealed a two-fold increase in pro/pre-B-I (fraction BC) cells, suggesting that the increase in CTG2 mice may simply be the result of the inability of two HCs to dimerise and escape the ER (Kaloff and Haas, 1995; Shimizu et al., 2002). It would be interesting to perform repertoire analysis on $SLC^{-/-}$ mice to find out if a similar skew in productive to unproductive recombination was present.

By fraction D, the ratio of productive to unproductive more closely resembles the expected WT ratio. However, the count of fraction D B cells is severely diminished compared to WT, as would be expected from the $SLC^{-/-}$ and $sIgM^{-/-}$ studies (Nguyen e al., 2015; Shimizu et al., 2002). Switching out the pre-BCR with an autoreactive antibody allowed developmental progression as well as proliferation, demonstrating the importance of autoreactivity at this stage of development (Köhler et al., 2008). The lower fraction D numbers suggests the paired HCs of CTG2 mice are able to signal developmental progression, but not proliferation. Repertoire data from fraction C'DE CTG2 B cells did not reveal any inclination towards autoreactivity, supporting the notion that the signalling initiated by the HCs is different or insufficient to induce proliferation.

The lack of proliferation during the autonomous receptor-signalling phase of the HCAb pre-BCR is reminiscent of the premature silencing of PI3K by SLP-65 (Figure 1-4). It is conceivable that the temporal separation of PI3K activity and the delay in SLP-65 function, suggested to allow proliferation in normal development (Herzog et al., 2009), is no longer present with a HCAb pre-BCR. Because the pre-BCR and BCR are conceivably identical at both stages of development (fraction BC and E), it is plausible that the signalling of the HCAb pre-BCR is indistinguishable from the BCR signalling and B cells are rushed through the intervening developmental stages. Analysis of the phosphorylation status of key signalling molecules downstream of the pre-BCR could reveal if the conventional pathways are utilised in the case of an HCAb pre-BCR (Figure 1-4).

The second accumulation of B cells was observed between the fraction E and fraction F stages of B cell development in the bone marrow. The majority of these intermediate cells did not express detectable levels of surface BCR, suggesting these cells constitute bone marrow resident B cells rather than peripheral migrants (Pelanda and Torres, 2012). B cells that fail to express the HC receptor on the cell surface can get through the pre-BCR checkpoint (Guloglu and Roman, 2006), but most likely fail to pass through BCR checkpoint based on the ubiquitous BCR expression on peripheral B cells. The signalling exhibited by the HC at the pre-BCR checkpoint is most likely carried over to the BCR checkpoint, whereby some form of tonic signalling maintains B cells and allows accumulation instead of deletion. The possibility of an autoreactive BCR triggering secondary recombination or V gene replacement that might keep a B cells in a prolonged loop due to the lack of a LC is unlikely. This is because cells undergoing these processes have been shown to acquire a de-differentiation phenotype and resemble earlier fraction B cells rather than an intermediate of more mature B cells (Schram et al., 2008; Tze et al., 2005). However, B cells with the intermediate E F phenotype, but lacking IgG surface expression could be explained by rapid endocytosis of the IgG BCR that has engaged with (auto)antigen (Knight et al., 1997; Song et al., 2016). The levels of autoantigens would be expected to be much higher in $sIgM^{-/-}$ mice due to the importance of IgM in cellular debris clearance (Ehrenstein and Notley, 2010), and might trigger, to a lesser extent, the BCR without the typical positively charged amino acid residues. Subsequently the immature B cells that have internalised their BCR are prevented from exiting the bone marrow niche and instead accumulate between immature and mature B cells. Repertoire or RNA sequencing of these intermediate B cells may shed more light onto their origin. In summary, the alterations in B cell number within the bone marrow developmental fractions is most likely a combination of impaired SLC binding, lack of secreted IgM and the signalling of an IgG-BCR that does not have a light chain.

The CTG2 mouse also displayed altered proportions of peripheral B cell in the spleen. Similar to $sIgM^{-/-}$ (Baker and Ehrenstein, 2002; Nguyen et al., 2015; Tsiantoulas et al., 2017), CTG2 mice displayed a

significant increase in MZ and a decrease in FO B cells. The same observations were also made in mice engineered to only express the IgG isoform throughout development (Waisman et al., 2007), but because these mice would have lacked a sIgM, the role of IgG-BCR in directing development of peripheral B cells is not conclusive. It would be interesting to transfer serum sIgM into IgG-BCR mice to examine if the two mechanisms can be untangled (Nguyen et al., 2015). In addition, $SLC^{-/-}$ mice also displayed a skew in the number of MZ to FO B cells, caused by a severe reduction in the number of FO B cells (Keenan et al., 2008; Ren et al., 2015). The B-lymphopenia observed in these $SLC^{-/-}$ studies does not seem to apply to CTG2 mice, where the size of the spleen as well as total $B220^{+}$ B cells was not significantly different from WT. It is thought that the escape of autoreactive BCRs from central and peripheral tolerance may be the cause of decreased FO and increased MZ B cells. However, CTG2 repertoire sequencing did not reveal a skew towards autoreactivity, suggesting that the negative selection that resulted in B-lymphopenia in $SLC^{-/-}$ was not in effect in CTG2 mice. Sequencing the repertoire of CTG2 MZ and FO B cells could reveal the additional role of peripheral tolerance and answer if autoreactive BCRs are observed in the periphery. Experiments monitoring the levels of Ca^{+} flux induced by the CTG2 HC receptors would be ideal to dissect signalling strength that biases development into MZ B cell subset. It is tempting to speculate that the lower fluorescence intensity of IgG observed in CTG2 mice compared to IgM in WT is an indication of the Ig surface levels and hence a proxy for signalling strength. However, the use of different antibodies with potentially different titrations prevents a fair comparison. In summary, the alteration of bone marrow B cell development and the selection of immature B cells with particular levels of BCR signalling plays an important role in guiding development into the different peripheral B cell subsets.

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

4 Building an analysis package for the unbiased quantification of recombination and optimising the VDJ-seq method

4.1 Background

The advent of next generation sequencing (NGS) has led to the adaptation and development of techniques for the high-throughput interrogation of antigen receptor (AgR) repertoires. The ability to sample thousands to millions of recombined sequences is enabling the examination of repertoire features, such as alteration of the repertoire during immune response (Hou et al., 2016a) or ageing (Martin et al., 2015), and answering fundamental questions about the regulation of recombination (Bolland et al., 2016). Understanding the factors regulating recombination at the chromatin level first requires an unbiased readout of V gene usage at the DNA level. For this purpose, we developed a DNA-based NGS technique, VDJ-seq, that quantitatively captures the products of recombination (Bolland et al., 2016; Chovanec et al., 2018; Matheson et al., 2017). The technique exploits the fact that each recombination contains one of a limited number of J genes, which can be captured after sonication with a single round of primer extension using a biotinylated J-specific oligo. The simple capture design of VDJ-seq overcomes the need for pairs of primers specific for V and J genes, which produce inherent biases (Baum et al., 2012), and enables it to be easily extended to any antigen receptor locus in any species with just a set of nested J gene primers.

One of the challenges in repertoire quantification is distinguishing true diversity from erroneous diversity created by PCR and sequencing errors. Each V(D)J-recombination event results in a unique sequence through combinatorial joining of the V(D)J gene segments, along with exonuclease nibbling and P- and N-nucleotide additions at the junctions between V-D and D-J genes. Additional diversity of sequences is produced by point mutations introduced by somatic hypermutation (SHM) during affinity maturation. Because of all these natural diversification mechanisms, a single nucleotide error could be incorrectly interpreted as a novel clone. The negative impact of PCR and Illumina sequencing errors has been extensively examined in RNA based repertoire quantification and methodologies to overcome them using unique molecular identifiers (UMIs) have been implemented (Bolotin et al., 2012, 2013; Shugay et al., 2014). UMIs usually consist of randomly generated sequences with high diversity, generally 8-12 bp in length, that are used to uniquely barcode individual DNA or RNA molecules (Fu et al., 2011; Greiff et al., 2015; Kivioja et al., 2012). Because of diversity overestimation through sequence errors, the importance of UMI use for AgR repertoire analysis has been highlighted in many recent studies. However, such approaches have been lacking for DNA-based repertoire quantification. To my knowledge VDJ-seq is the first DNA-based repertoire sequencing method to take advantage of UMIs (Wardemann and Busse, 2017).

VDJ-seq can quantitatively capture both productive and unproductive VDJ, DJ and even abnormal recombination events. This makes it a valuable tool for the study of the primary output of V(D)J recombination and repertoire selection in wildtype and knockout models. Most assays developed for profiling AgR repertoires utilise RNA instead of DNA and therefore are only able to capture productive and not yet degraded unproductive VDJ recombination. The higher abundance of RNA of the same recombination event within a single cell results in RNA-based assays having better sensitivity. The widespread use of RNA has also resulted in the creation of dedicated analysis tools that take advantage of the multi-copy nature of RNA to correct early PCR errors in addition to the more frequently observed PCR and sequencing errors (Chaudhary and Wesemann, 2018; Shugay et al., 2014). However, DNA-based approaches also have additional advantages over RNA. The RNA output from a recombination event varies by orders of magnitude between cells. The different activation states of a cell, differences in mRNA stability along with V gene promoter strength, which stems from the evolutionary V gene family of origin, all contribute to the varying levels of RNA (Bolland et al., 2016; Choi et al., 2013a; Love et al., 2000; Wardemann and Busse, 2017). As a result, a single contaminating antibody secreting B cell or a strong promoter can easily skew or mask the actual repertoire. It is for this reason that most of the RNA-based repertoire analyses are confined to groups of identical CDR3 and VJ gene sequences (clonotypes). In contrast, each productive recombination event is present at only a single DNA copy within a cell, allowing VDJ-seq to provide a true picture of clone numbers within the captured repertoire.

Despite the many advantages of VDJ-seq (Bolland et al., 2016; Chovanec et al., 2018; Matheson et al., 2017), there are several limitations that are worth noting. First, as with all DNA-based repertoire sequencing methods, the linear distance between the VDJs and the constant region exons prohibits their amplification and sequencing and therefore prevent the assignment of isotype to the captured VDJ recombination events. Second, the capture of target sequences using J oligos means that all unrecombined germline J DNA fragments will also be pulled-out and sequenced. Thus, a large portion of the library will contain uninformative reads. The initial VDJ-seq method included a depletion step where a second primer extension step, with biotin oligos positioned in the intronic regions between J genes, was used to remove germline DNA fragments. Despite this depletion step the final libraries still contained a large number of germline reads (Bolland et al., 2016; Matheson et al., 2017). Nevertheless, the germline reads proved essential for the implementation of mispriming correction in the analysis pipeline (see section 4.4) and together with VDJ and DJ recombination they provide a complete status of recombination within a population of cells (see section 4.3), which is important in studies focusing on impaired recombination. Third, a limitation of all bulk-population AgR-seq methods is the inability to capture the pairing information of TCRα-TCRβ, TCRγ-TCRδ and of the Igh

heavy-light chains. To address this shortfall, single-cell methods have been developed that either merge the mRNA transcripts from the two loci using overlap extension PCR (DeKosky et al., 2013; McDaniel et al., 2016) or generate cDNA with template switch oligos in droplets containing cell barcodes, which allow the two chains to be linked *in-silico* (10x Genomics). Although single-cell methods are likely to be the future of repertoire profiling, the current implementations still require specialised equipment, high capital investment and substantial consumables costs that prohibit their widespread use (Dunn-Walters et al., 2018). A final limitation of VDJ-seq stems from its use of sonication to derive a set of fragments that is below the 1 kb limit of efficient cluster formation of Illumina sequencers. An optimal size of 500 bp was chosen for the Igh, while 400 bp was enough for the Igκ due to its smaller recombinant size. The size allows the 300 bp paired-end chemistry to sequence the full length of fragments. However, the random nature of sonication results in the majority of captured VDJ sequences to not contain a full-length V gene sequence. This does not pose a problem for obtaining a complete CDR3 and assembling clonotypes, but it can limit the characterisation of somatic hypermutation within the CDR1 and CDR2 regions of the V gene. Despite these limitations, VDJ-seq is a powerful new tool for examining antigen receptor repertoires, harnessing the advantages that come with using gDNA as the starting material.

4.2 Aims

The current VDJ-seq protocol is limited to large quantities of material, preventing the analysis of rare populations of B cells. Existing repertoire sequencing protocols have highlighted the prevalence of PCR and sequencing errors that can cause artificial inflation of a repertoire's diversity and have demonstrated the importance of UMIs in tackling this challenge. In addition, UMIs provide a simple molecule counting method through which true biological duplicates can be distinguished from PCR duplicates. The lack of a UMI or the presence of an insufficiently diverse UMI have limited the quantitative power of VDJ-seq.

Additionally, the current analysis pipeline does not address the issues of PCR and sequencing errors and mispriming observed between all the J gene primers. The removal of PCR duplicates is the only current use of UMI sequences, which underutilises their additional ability to discriminate chimeric and artefactual reads from true recombination events and obtain a high-quality consensus sequence.

To address the current limitations of VDJ-seq and the analysis pipeline I sought to:

1. Optimise VDJ-seq for low input samples with a sufficiently diverse UMI and establish required levels of over-sequencing that would enable error correction.
2. Construct a comprehensive analysis pipeline that enables error correction, artefact sequence removal and accurate quantification of clonal sequences.

4.3 The optimisation of VDJ-seq

The initial VDJ-seq was established using large amount of starting DNA (5-10µg) that meant it was limited to samples from which high numbers of B cells could be enriched. The ability of VDJ-seq to accurately quantify clone numbers makes it ideally suited for studies of clonal expansion, immune dysfunction and a host of other potential clinical applications (Georgiou et al., 2014). But to profile the limited quantities of material usually obtained from clinically relevant samples, I first needed to establish an optimised VDJ-seq protocol that produces reproducible results from low starting material. I took a systematic approach to optimise each step of the protocol were possible (Figure 4-1).

The first change to the protocol was the use of the New England Biolabs (NEB) end-repair, A-tailing and ligation mix that allows all three reactions to be performed sequentially within a single tube (Figure 4-1). This removed two clean-up steps during which it is common to lose around 10-20 % of material. The purification after adaptor ligation was switched from QIAquick PCR purification columns to AMPure XP beads as the magnetic beads were able to remove higher amounts of unligated adaptors. To test the impact of different polymerases on the final VDJ-seq library I performed experiments with three polymerase combinations using the same starting material. One of the combinations involved switching the primer extension polymerase Vent (exo-) for the Q5 high-fidelity polymerase. The capture efficiency, quantified as the number of final VDJ reads divided by a theoretical total based on the starting material, was lower when the Q5 polymerase was used for both primer extension and the two rounds of PCR (Q5/Q5 1 µg) compared to the combination with Vent (exo-) (Vent/Q5 1 µg) (Figure 4-2 a). Despite the lower capture efficiency, the Q5 polymerase did produce reads with fewer errors (Figure 4-2 b). But, as these can be corrected using the UMI groups, it is better to have high capture efficiency by sticking with the Vent (exo-) polymerase for primer extension and using the Q5 for the two round PCR (Figure 4-2 a). With increasing starting material, the capture efficiency also seems to improve, reaching around 10% for 10 µg (data not shown). To further limit the loss of material I removed the steps used for depleting unrecombined fragments using the primer extension of the germline J gene regions with biotinylated primers. I reasoned that based on the efficiency of the primer extension capture and the dominance of germline reads in libraries prepared with a depletion step, only a small portion of germline sequences are being effectively removed, while the additional steps result in loss of material that could be detrimental to low material samples. The retention of germline reads also allows the analysis of the complete recombination status of a population of cells without the skew of germline depletion. Finally, a 12 N UMI adaptor was introduced to replace the previous 6 N adaptor, which I found out contained insufficient diversity for it to be solely used for grouping and deduplicating reads. Overall, these optimisations of the VDJ-seq

protocol allowed library generation from as little as 500 ng starting material, which is approximately equivalent to 100,000 cells.

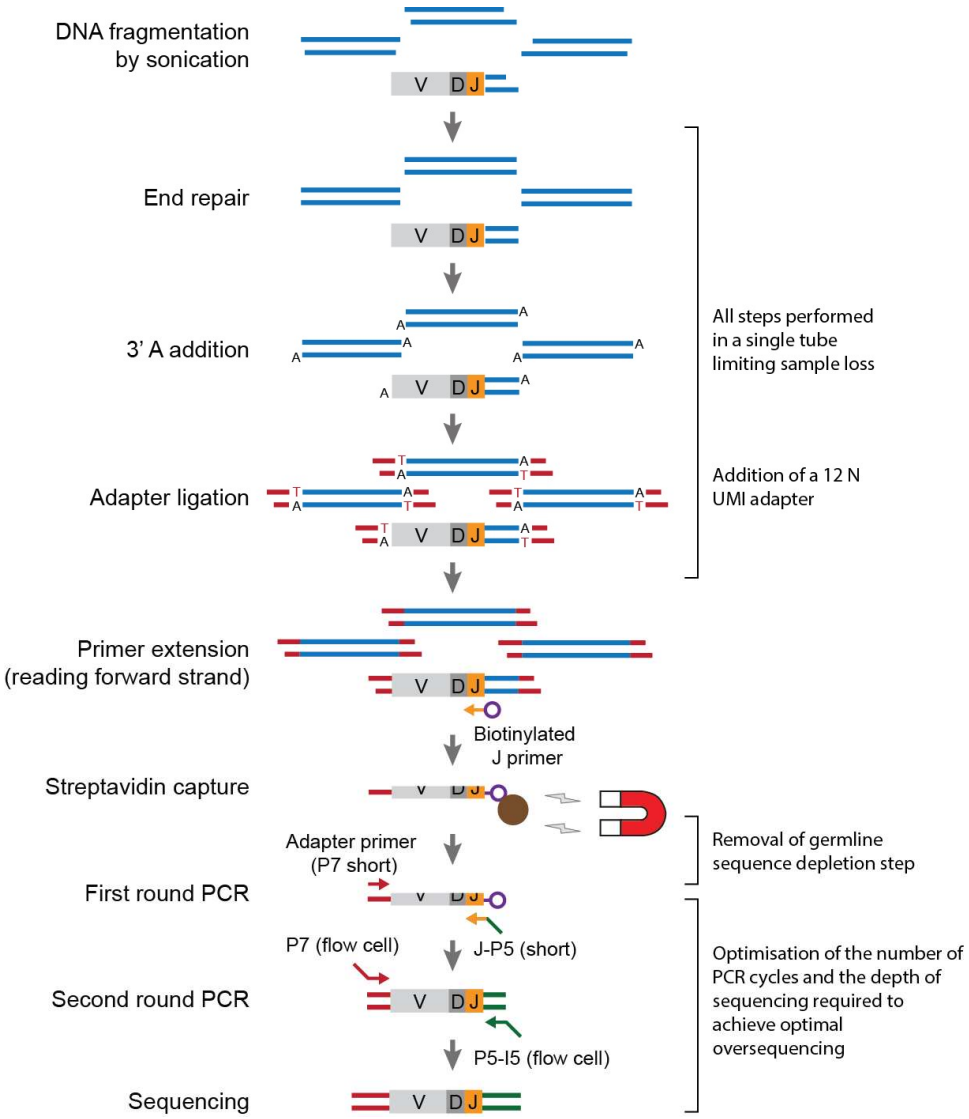


Figure 4-1: Optimised VDJ-seq assay for low starting material.
For low starting material samples, the end repair, A-tailing and adaptor ligation are performed in a single tube to reduce material loss. A new longer (12 N) UMI is used. The depletion of germline reads performed previously (Bolland et al., 2016; Matheson et al., 2017) has been removed, as the additional steps lead to material loss which outweighs the gain in final library purity. In addition, the number of PCR cycles along with the required sequencing depth have been optimised to produce a library with enough over-sequencing that would enable confident PCR and sequencing error correction.

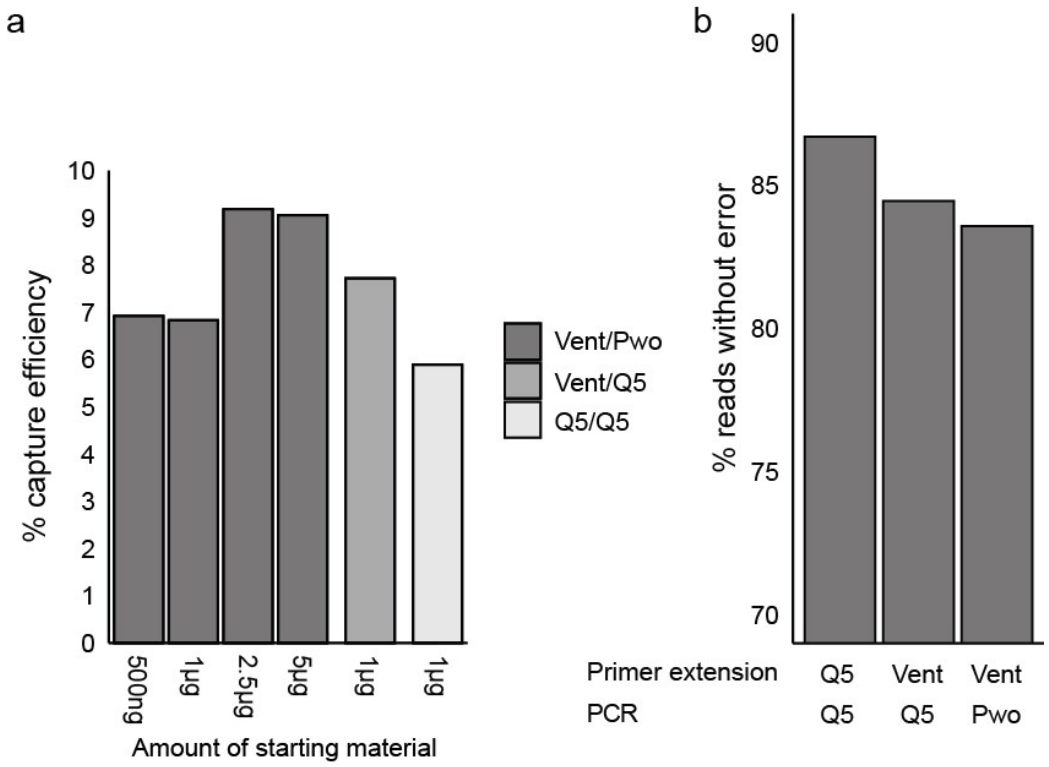


Figure 4-2: The choice of polymerases impacts capture efficiency and the amount of errors within reads.
(a) With increasing starting material, the capture efficiency of VDJ-seq sequences also increases. The capture efficiency was calculated as the amount of final VDJ products captured over the total theoretical sequences expected base on the starting material. The use of the Vent (exo-) polymerase for primer extension along with the Q5 high-fidelity polymerase (Vent/Q5) produced a 1% higher capture efficiency compared to the Vent (exo-) and Pwo polymerase combination (Vent/Pwo). The use of the Q5 polymerase for primer extension is less efficient than Vent (exo-) as illustrated by the almost 2% decrease in capture efficiency when the Q5 polymerase is use for both primer extension and the subsequent PCRs (Q5/Q5). (b) The number of mismatches between sequences of a UMI group and the consensus sequence of that group can be used as an estimate of the errors generated during the library preparation process. The combination of the Q5 polymerase for both primer extension and PCR results in the lowest number of errors within reads, while the Vent (exo-) and Pwo polymerase combination produces the most within read errors.

To establish the versatility of VDJ-seq for different amounts of starting DNA, I generated libraries from 500 ng, 1 µg, 2.5 µg and 5 µg. From previously made VDJ-seq libraries, we roughly knew the required amount of amplification needed to generate an optimal library concentration (=> 2 nM) from samples in range of 1 µg - 10 µg. To ensure the level of sequencing was enough to allow the correction of PCR and sequencing errors, I first examined the duplication level within the four libraries (Figure 4-3 a). The duplicate distribution for each library showed a clear peak at around 10-15 reads per UMI group, with just the left most tail falling off the plot. This reassured me that the majority of UMI groups contained enough reads. I next extrapolated the level of additional unique UMIs expected to be observed with more sequencing depth using species accumulation curves from the preseqR package (Deng et al., 2015) (Figure 4-3 b). The extrapolation (Figure 4-3 b - dashed line) demonstrated that the current sequencing depth (Figure 4-3 b - solid line) already surpassed the inflection point of the accumulation curve and therefore shows that only marginal gains in new UMIs would be obtained with more sequencing.

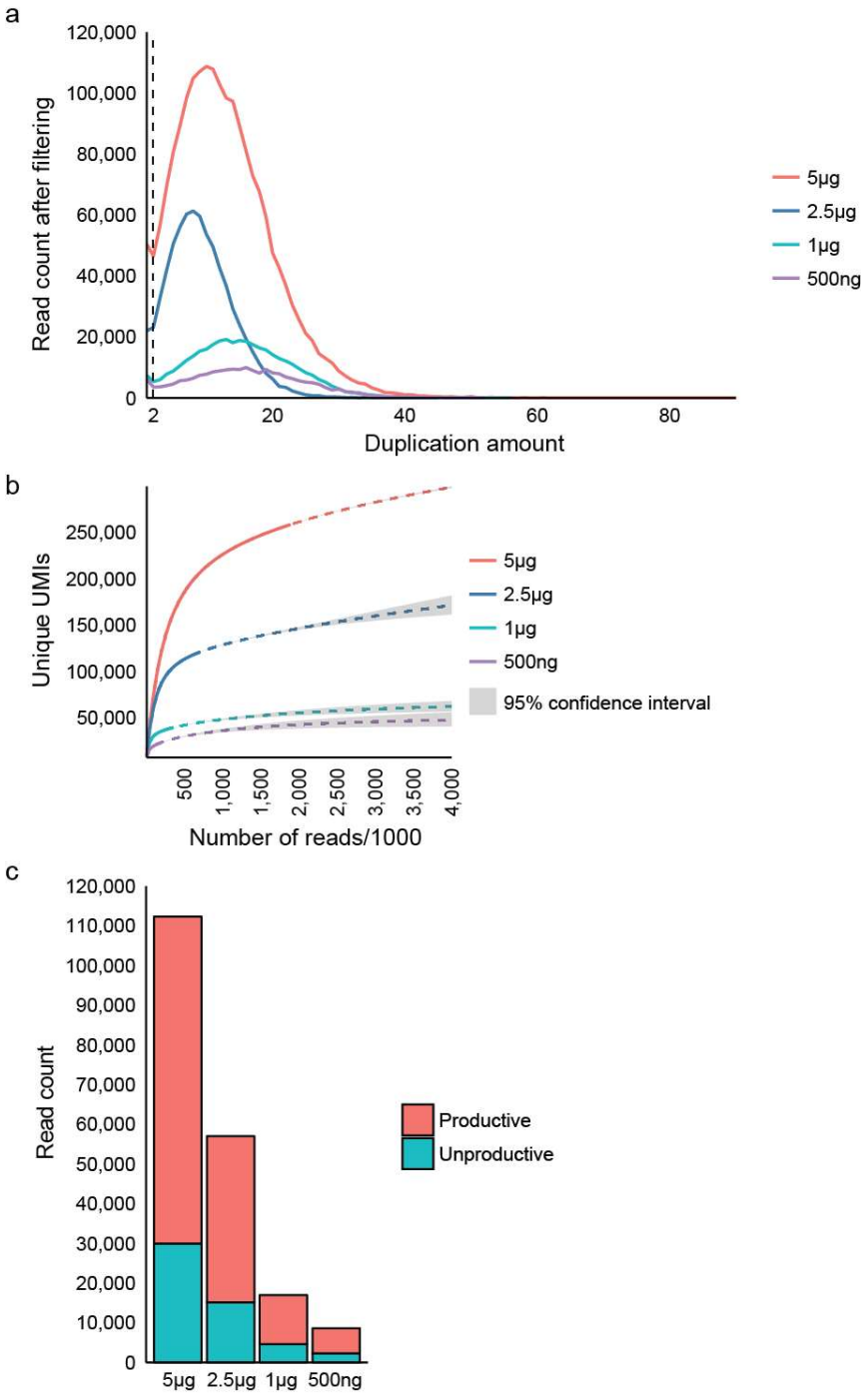


Figure 4-3: A range of starting materials demonstrate the versatility of VDJ-seq.
(a) Levels of duplicates present in VDJ-seq libraries made from different amounts of starting material. UMI groups with high number of sequences are desired, as they allow the high confidence correction of PCR and sequencing errors. The read counts shown are after removing sequences that diverged from the consensus sequence (good-to-total ratio filter). The vertical dashed line shows the reads that will be filtered by the low UMI read count cutoff (UMI group with two or fewer reads will be removed). (b) To examine the levels of over-sequenceing present within the different starting material libraries, the preseqR (Deng et al., 2015) package was used to estimate the number of additional unique UMIs that would be observed with more sequencing (dashed line), based on the observed UMIs (solid line). (c) Number of final VDJ sequences in spleen B cells obtained from different quantities of starting material. The number of VDJ sequences will vary depending on the population of B cell sequenced.

The expected ratio for spleen B cells of 2:1 productive to unproductive recombination events is obtained from all four libraries, with the 500 ng starting material library still capturing nearly 10,000

VDJ recombination events (Figure 4-3 c). The ratio, along with the high reproducibility of V gene usage between high and low amounts of starting material (5 µg vs 500 ng; r^2 = 0.989), suggests that low DNA concentrations do not introduce noticeable biases into the final VDJ-seq library (Figure 4-4 a). I did not make any libraries with less than 500 ng starting material, as I reasoned that for 20000 cells (100ng), based on a 7% capture efficiency, one would expect to capture 1400 VDJ sequences, 933 productive, which given the large dynamics range of recombination frequencies observed within a diverse repertoire (200 fold for mouse Igh) would result in dropouts, making V gene usage statistics possible only for highly recombining V genes. However, there is nothing in the VDJ-seq protocol that would prevent starting from lower amounts (100 ng) with an increased number of PCR cycles, ultimately making it possible to analyse rare populations such as plasma cells.

The relatively low capture efficiency of VDJ-seq and the high diversity of antibody repertoires results in a very low clonotype overlap between technical replicates (Figure 4-4 b). This shows that a different portion of the vast repertoire is captured within each technical replicate. Even samples made using high amounts of starting material share very few clonotypes (a 0.45% overlap between 5 µg and 2.5 µg samples). This opens the possibility of performing multiple technical replicates as a way of recovery a large portion of the total repertoire.

4.3.1 VDJ-seq outperforms other DNA based repertoire sequencing methods

Besides VDJ-seq, there are two other repertoire sequencing methods that use DNA instead of RNA for the capture of recombination. The initial studies of repertoire were performed using a large set of V and J primer pairs. These methods were later adapted for use with next generation sequencers, allowing the examination of thousands of sequences at once (Georgiou et al., 2014). To compare VDJ-seq with these earlier methods, I started off by examining the functional V gene usage quantified by VDJ-seq and a VJ primer based method (Kaplinsky et al., 2014). The comparison revealed a high number of differences (Figure 4-5 a and Figure 4-6 a). Several genes detected by VDJ-seq were absent from the VJ primer dataset, while other that were highly recombining according to VDJ-seq were only lowly detected using VJ primers. This is perhaps not surprising given the varied amplification efficiency of each primer pair and possible primer cross-reaction (Kaplinsky et al., 2014; Wardemann and Busse, 2017). A more recent DNA-based technique that has adapted a genome-wide translocation sequencing assay (LAM-HTGTS) (Hu et al., 2016) for repertoire sequencing (HTGTS-Rep-seq, same principle of J gene primers with the addition of linear amplification and single stranded adaptor ligation) (Lin et al., 2016) shows high agreement in V gene usage with VDJ-seq (Figure 4-5 b and Figure 4-6 b). However, the absence of UMIs or a deduplication step in the HTGTS-Rep-seq analysis pipeline resulted in a high number of identical sequences in the final dataset. I attributed these sequences to be PCR duplicates as naïve spleen B cells do not tend to have high clonality (Silva and Klein, 2015).

Because I could not definitively exclude the possibility that they represent clonal sequences, I chose to perform a clonotype comparison between VDJ-seq and HTGTS-Rep-seq. The number of clonotype groups obtained from comparable amounts of starting material was 2-5 times higher in VDJ-seq compared to HTGTS-Rep-seq, demonstrating that VDJ-seq has a greater capture efficiency than HTGTS-Rep-seq (Figure 4-6 c). In summary, these comparisons highlight some of the advantages of the unbiased repertoire capture produced with VDJ-seq, compared to other DNA-based repertoire sequencing methods.

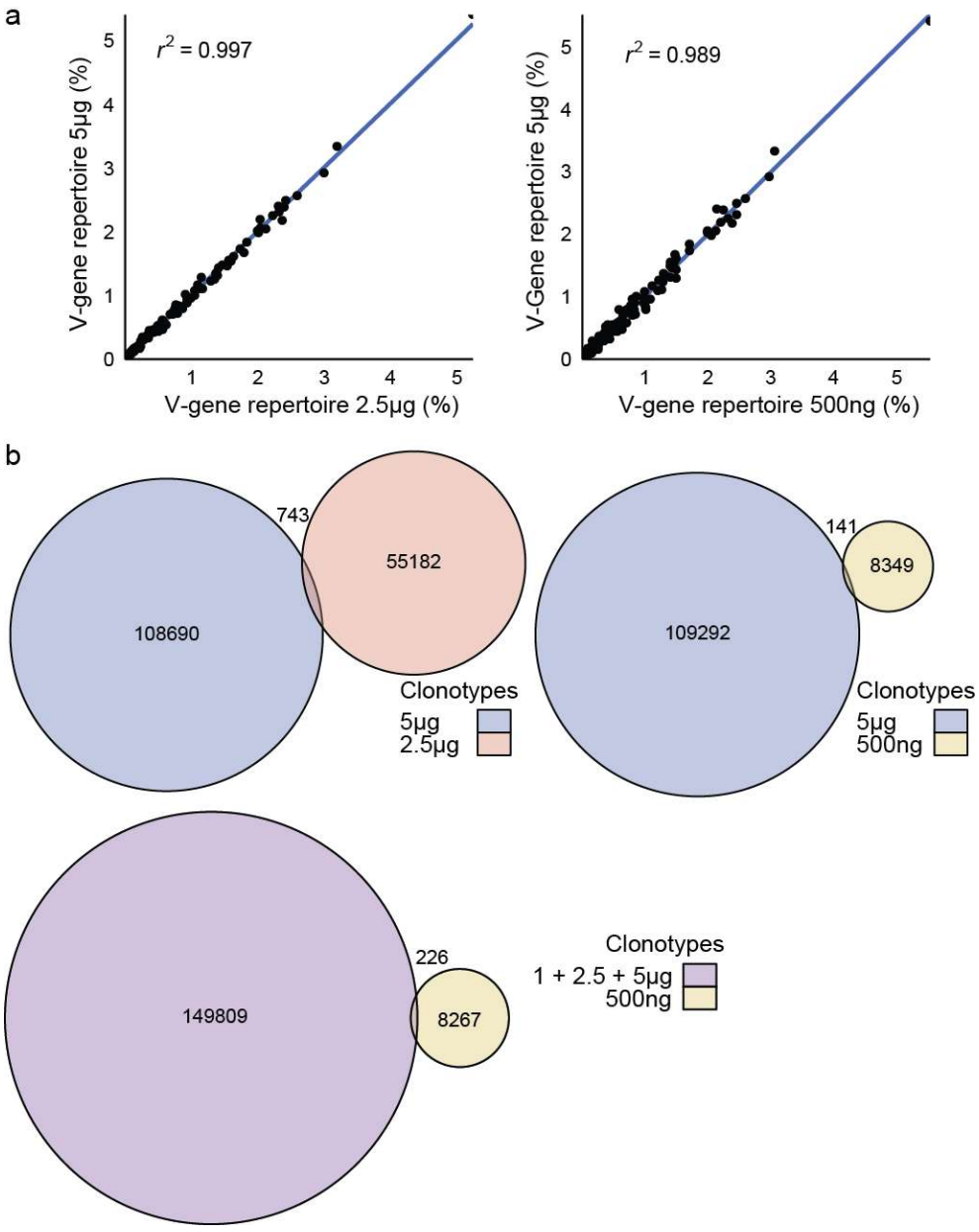


Figure 4-4: VDJ-seq produces highly reproducible libraries from different amounts of starting material.
Data was generated using B cells enriched from two pooled 12-week-old C57BL/6 female mouse spleens. (a) The proportion of each V gene observed within VDJ-seq libraries is consistent between low and high starting material samples. (b) Venn diagrams depicting clonotype overlap between libraries prepared from different amounts of starting material. The clonotypes here are defined as VDJ recombinations that share the same VJ gene annotation and have an identical CDR3 nucleotide sequence.

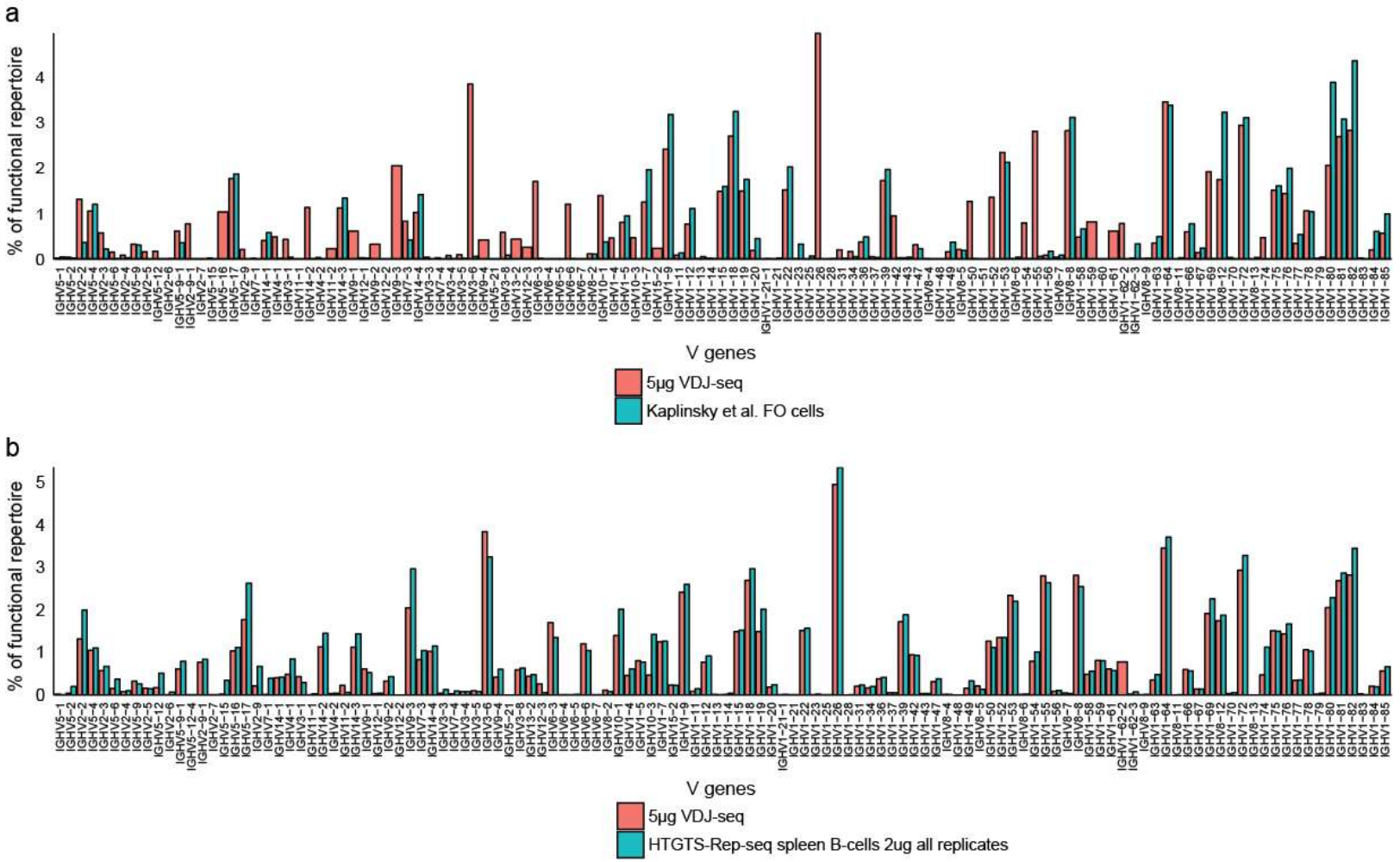


Figure 4-5: V gene usage of spleen B cells observed with VDJ-seq and two other DNA-based methods.
The thicker bars highlight V genes missing in one of the samples. (a) Repertoire data generated from follicular B cells using a cocktail of VJ primers (Kaplinsky et al., 2014) compared to data generated with VDJ-seq from total spleen B cells. The VJ primers fail to capture several V genes across the mouse Igh locus. (b) Repertoire data generated from 2 µg of starting material extracted from spleen B cells using the HTGTS-Rep-seq method (Lin et al., 2016) compared with 5 µg startin material VDJ-seq library. The absence of the Ighv1-62-2 gene from the HTGTS-Rep-seq sample is likely due to it sharing an identical sequence with the Ighv1-71 gene. Filtering of multi-V gene calls would have removed it from the results.

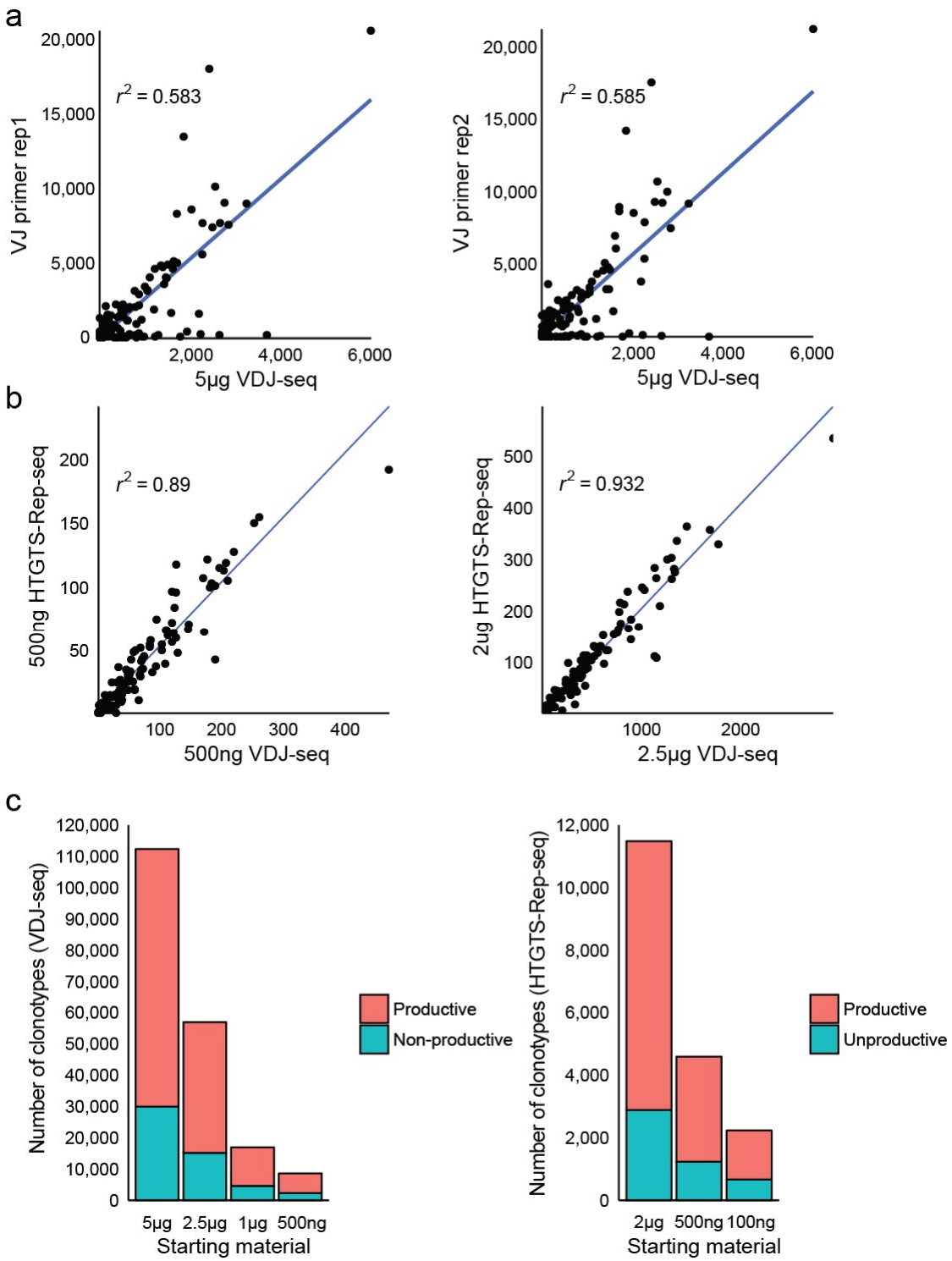


Figure 4-6: Further comparison of VDJ-seq to other DNA-based repertoire sequencing methods.
(a) The proportions of V genes captured with the VJ primer cocktail method have a low agreement with VDJ-seq. (b) Unlike the VJ primer cocktail, VDJ-seq and HTGTS-Rep-seq have high agreement of V gene usage when comparing libraries made from similar starting material. (c) HTGTS-Rep-seq data contains many identical sequences, which I interpreted as PCR duplicates as naïve spleen B cells are not expected to have high clonality. As a result, I chose to compare the number of observed clonotypes captured with each assay. VDJ-seq with equivalent starting material captures a much higher number clonotypes compared to HTGTS-Rep-seq.

4.4 The BabrahamLinkON pipeline for the analysis of VDJ-seq data

The herein described BabrahamLinkON pipeline (available on: https://github.com/peterch405/BabrahamLinkON) builds upon earlier work done by Felix Krueger (Bolland et al., 2016; Matheson et al., 2017) with contributions from Dan Bolland, Louise Matheson, Bryony Stubbs and Amanda Baizan-Edge. The initial pipeline was designed for mouse Igh and Igκ recombination. Mispriming correction was limited to the most frequent mispriming event between J2-J4 in the Igh, while for the Igκ a more extensive find and replace strategy was used. Subsequently, deduplication was performed based on the unique sequence 60bp downstream of the J primer and the V read start position for the Igh. Because of the lower diversity of the Igκ locus, an additional 6bp UMI with two anchor sequences were implemented (Matheson et al., 2017). The deduplication of Igκ reads uses the UMI with a single bp from the barcode, along with 20 bp from the J end read. The annotation of V segments was performed by bowtie (Langmead et al., 2009) alignment of V end reads to the mouse reference genome or with IMGT/HighV-QUEST (Alamyar et al., 2012). Despite the well-suited nature of the pipeline for V gene quantification, the desire to perform more in-depth quantitative analysis, such as clone analysis and assembly of clones into clonotypes, raised the need for a new pipeline.

The new pipeline divides the analysis of VDJ-seq data into three main stages: precleaning (orange), deduplication (green) and annotation with clone assembly (blue) (Figure 4-7). Aspects of each stage of the pipeline are described in more detail in the sections to follow. The precleaning stage of the pipeline first attempts to assemble the paired-end reads using the Paired-End reAd mergeR tool (PEAR) (Zhang et al., 2014). For reads obtained from 300bp paired-end sequencing runs, the assembly of reads tends to be above 95% (Figure 4-7). Germline sequence, sequences with low quality bases (Phred Q2; Phred is a measure of quality used in DNA sequencing; Q2, compared to other scores that predict an error rate, is an indicator that a certain portion of the read should not be used in further analysis) or sequences with UMIs containing bases with Phred quality scores of less than Q30 (a 1 in 1000 probability that the base is incorrect; 99.9% accuracy) are all filtered out. The majority of a VDJ-seq library will consists of germline reads (63.4%; Figure 4-7). Subsequently, J gene mispriming correction is performed and the J identity, along with the extracted UMI and anchor sequence are placed in the read name of each sequence in the output fastq file (Figure 4-8). The extracted UMI and anchor sequence is trimmed away as it is no longer needed within the read sequence (Figure 4-10 a). The anchor sequence is designed to secure the 3' end of the P7 adaptor oligonucleotide, allowing the UMI random nucleotides to form a bubble between the two stretches of double-stranded DNA. This allows the P7 adaptors to be efficiently ligated to A-tailed DNA. A pair of anchors is used to ensure a signal is generated in both the red (A/C nucleotides) and the green (G/T nucleotides) laser channels of

the Illumina sequencers. Based on the pair of anchors, the sequences are split into two groups that are deduplicated separately of each other. After precleaning, around 20% of the starting reads are left and ready for deduplication (Figure 4-7).

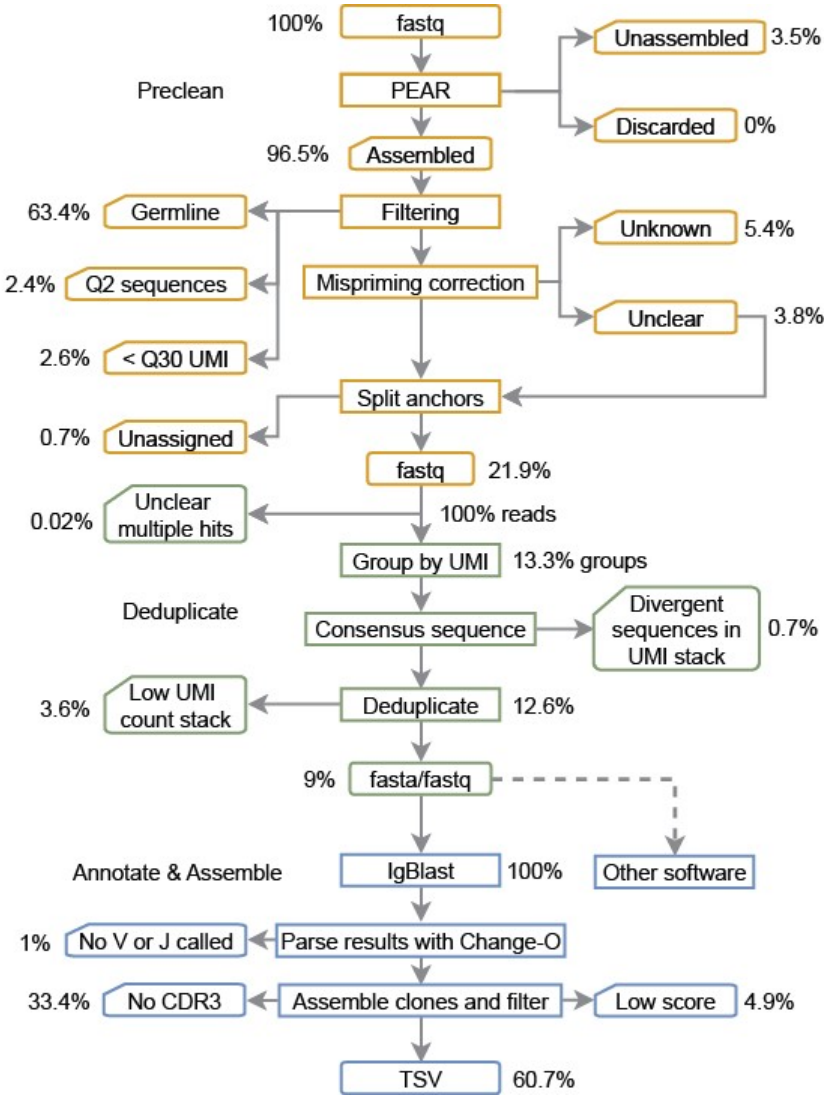


Figure 4-7: BabrahamLinkON analysis pipeline overview.
The pipeline is divided into three main stages. The first step involves assembly of paired-end reads into a single contiguous read, cleaning the data and correcting mispriming events (orange: precleaning). The UMI is also extracted from the reads at this stage and placed into the read name. Next, the reads are grouped according to their UMI sequence and a consensus sequence is derived. A set of criteria such as divergence of sequences in a UMI stack also filters incorrectly grouped sequences, chimeric reads or reads with substantial PCR and sequencing errors (green: deduplicate). Finally, the VDJ and DJ reads are annotated using IgBlast (blue: annotate and assemble). There are many tools that perform annotation and clone assembly and the output fasta/fastq from the deduplication stage can be used with any of them. The read percentages at each stage are based on a 5µg starting material sample from B cells enriched from the spleen of a female 12-week-old C57BL/6 mouse. The annotation was not performed for DJ recombination events, meaning most of the reads being removed at the annotation and assembly stage represent DJ reads.

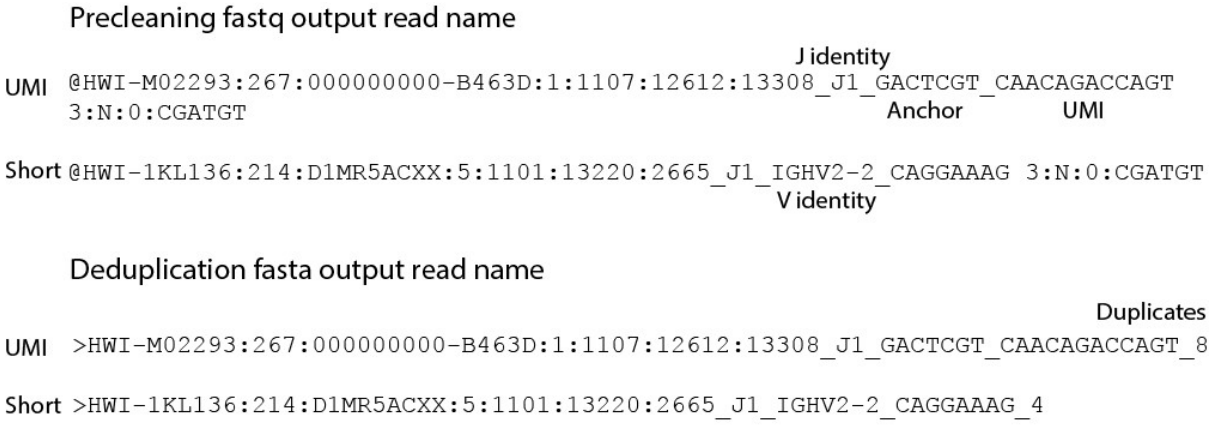


Figure 4-8: Information deposited into the read name during different stages of the pipeline.
The information required for deduplication is extracted at the precleaning stage and placed into the read name. This includes the anchor sequence and the UMI. In addition, the J gene identity from mispriming correction is included and can be used to verify J gene annotation in the last stage of the pipeline. For short reads the V gene identity based on Bowtie 2 alignment is also included. At the deduplication stage, the number of duplicates for each output read is written into the read name. This gives the user the flexibility to filter reads later based on their duplicate count.

The majority of reads within a VDJ-seq library are duplicate sequences (86.7%) of the base (13.3%) reads that form the foundation of each unique UMI group (Figure 4-7). After bundling the reads with identical UMI sequences, a multiple sequence alignment (MSA) is performed using Kalign2 (Lassmann et al., 2009). The MSA allows a consensus sequence to be derived for each UMI group, forming the basis of the good-to-total ratio (gt-ratio) filter that was inspired by the MiGEC tool (Shugay et al., 2014). The gt-ratio works by first determining the hamming distance of each sequence within the UMI group compared to the consensus sequence. The hamming distance represents the number of positions in which two strings of equal length differ (mismatches). Sequences with a hamming distance of less than 5 (default setting) are marked as good sequences and a ratio of good to total sequences (by default 1; all sequences need to be less than 5 from consensus) for each UMI group is used to filter out ones with divergent sequences (Figure 4-7). A minimum of three sequences within each UMI group is required to reach a consensus, with each additional sequence increasing the consensus confidence. For this reason, over-sequencing of VDJ-seq libraries is desired. UMI groups with low sequence counts may be the result of errors within the UMI sequence and therefore low count UMI groups are also filtered out into a separate file (Figure 4-7). The same strategy has been previously used in RNA based repertoire sequencing methods where early PCR errors and errors introduced during cDNA synthesis result in UMI groups that cannot be confirmed by another independent RNA, resulting in so-called clonotypes singletons (Shugay et al., 2014; Turchaninova et al., 2016; Yaari and Kleinstein, 2015). Depending on the sequencing depth and library diversity of a sample, an optimal sequencing depth may not always be achieved. The read count of each UMI group is written into the read name of each output consensus sequence (Figure 4-8). This allows more stringent downstream filtering, in cases of enough over-sequencing, or the inclusion of low count UMI groups in cases of insufficient over-

sequencing. After deduplication, around 9% of error corrected and deduplicated reads are written out in either fasta or fastq format, ready for annotation and clonetype assembly (Figure 4-7).

The final stage of the BabrahamLinkON pipeline is performed using IgBlast v1.7.0+ (Ye et al., 2013) with the IMGT (internation ImMunoGeneTics information system) database and the Change-O v0.3.7+ (Gupta et al., 2015) IgBlast output parser for annotation. However, the output of the deduplication stage can be used with any of the publicly available annotation tools (Alamyar et al., 2014; Bolotin et al., 2015; Gaëta et al., 2007; Munshaw and Kepler, 2010; Ralph and Matsen, 2016; Ye et al., 2013). In addition to the IMGT reference, a custom D gene germline reference has been constructed that allows annotation of DJ recombination events. The reference is based on the human DJ reference concept first introduced by Duncan K. Ralph for the Partis annotation tool (Ralph and Matsen, 2016). The idea is to use the D germline sequence as a pseudo V gene that allows its seamless use in tools solely designed for VDJ annotation. After the IgBlast results are parsed with Change-O, sequences without or with a low score V or J call are removed along with sequences without a CDR3. Each sequence is assumed to arise from an individual cell and the expansion of each cell during B cell development gives rise to a pool of clones that all contain the same recombined sequence. At this stage of the pipeline, the clones are assembled together into a clonotype using their V and J gene annotation along with their CDR3 length and nucleotide composition. A single nucleotide mismatch between the CDR3 of clones that will be grouped/assembled is allowed by default. The clone assembly assigns a unique tag to each clonotype group, preserving the IgBlast annotation of each sequence/clone. This contrasts with RNA based tools, which tend to output only clonotype groups, as assigning an individual sequence to a single clone cannot be made due to the multi-copy nature of RNA. After annotation and clonotype assembly, around 60% of sequences are annotated as VDJ and are returned in a tab-separated file (Figure 4-7).

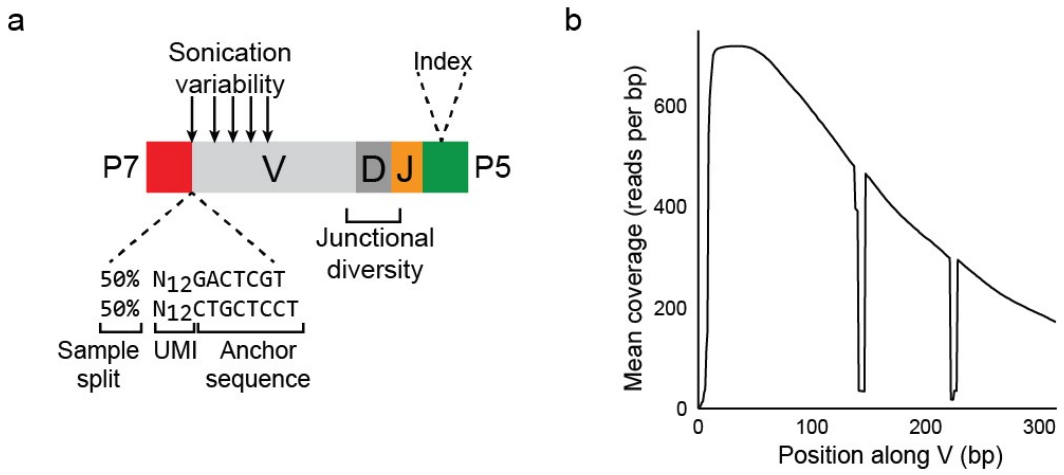


Figure 4-9: The composition of a DNA fragment in a VDJ-seq library.
(a) The final library contains DNA fragments with a P7 and P5 adaptors required for sequencing. In the P7 end, 50% of the library contains one of the two anchor sequences along with a 12N (random nucleotide) UMI. The P5 end contains the Illumina index for multiplexing samples. In samples prepared without an anchor-UMI it is possible to use the sonication variability in the V region along with the junctional diversity as a proxy UMI. (b) A mean coverage plot over the V region illustrates the variable V length produced by sonication.

The BabrahamLinkON pipeline is designed to accommodate four different forms of input data, depending on library preparation and sequencing length used. The latest generated libraries use a 12N UMI and are sequenced using long 300 bp paired-end sequencing runs (Figure 4-9 a). The long reads allow a straightforward read assembly and deduplication that is solely based on the diverse UMI sequence (the `umi` pipeline; Figure 4-10 a). However, extensive datasets have been generated within the lab using several different iterations of the VDJ-seq method. To be able to make use of all this data, three additional pipeline options have been created (Figure 4-10). I designed the first of these options (`short`) to accommodate short reads that cannot be assembled and that do not contain an anchor-UMI sequence (Figure 4-9 b). These types of samples represent the initial datasets generated with VDJ-seq (Bolland et al., 2016). Due to the lack of overlapping sequence, the primary analysis focuses on the J end reads. The V end reads are primarily used for their sonication variability proxy UMI and the more confident V gene annotation that can be attained from them. My initial analysis of this data using the gt-ratio filter revealed that the sonication variability alone contained insufficient diversity. Much like the previous pipeline, I chose to include a stretch of basepairs that roughly overlapped one of the highly diverse junctions form the J end read (Figure 4-10 b). In later iterations of VDJ-seq, a 6 N UMI with anchor sequences was added (Matheson et al., 2017), which required a separate pipeline option (`short_anchor`) (Figure 4-10 c). Although 6 N alone was still not diverse enough for some samples, the additional inclusion of the sonication variability proved to create a sufficiently diverse UMI for many UMI groups to pass the gt-ratio filter. The final pipeline option (`no_anchor`) is to accommodate long reads without an anchor-UMI (Figure 4-10 d). This option is only available in the deduplication stage of the pipeline and is a cross between the long reads in the

umi pipeline option and the proxy UMIs from the short pipeline option. Altogether, the BabrahamLinkON pipeline can handle a diverse range of starting sequences and use cases depending on the experimental design.

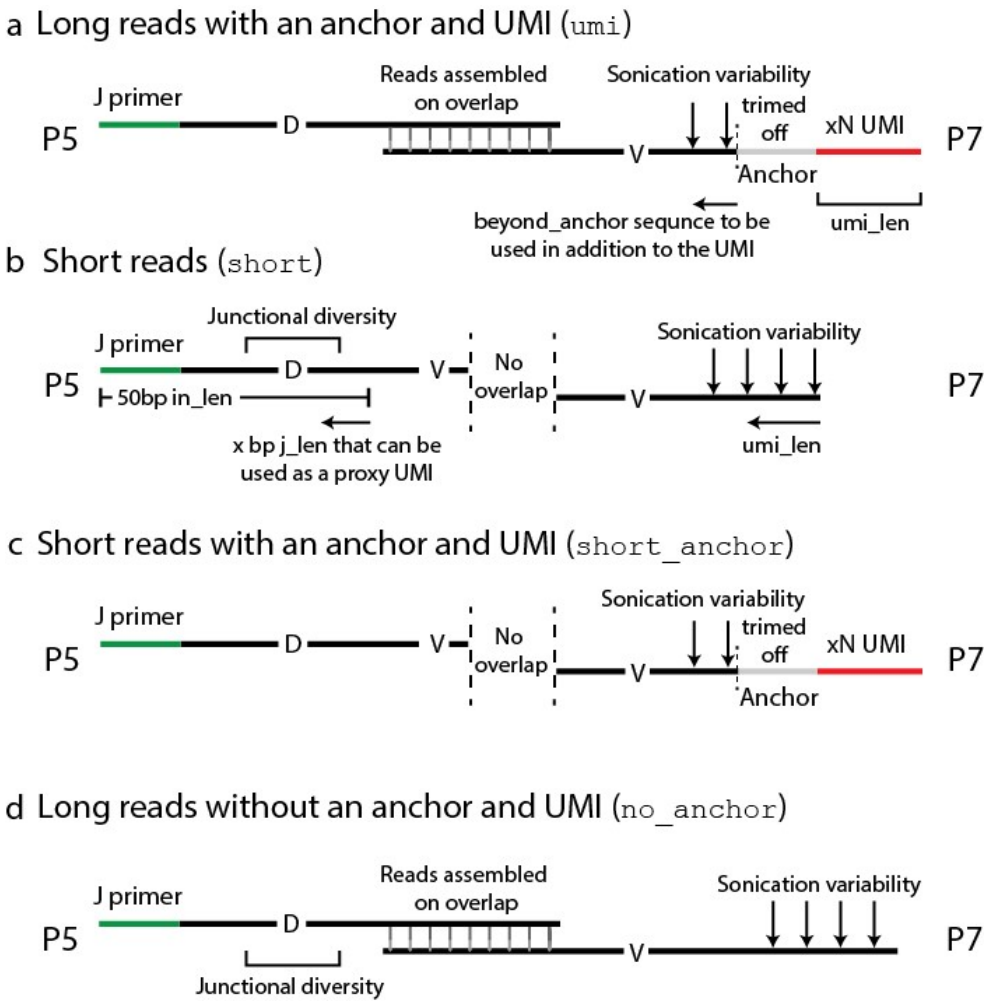


Figure 4-10: The four different use cases of the BabrahamLinkON pipeline.
(a) The latest and recommended version of VDJ-seq, which always contains at least a 12 N UMI and sequenced using longer than 250bp paired-end sequencing, should always be used with the umi option of the BabrahamLinkON pipeline. This option exclusively uses the UMI for deduplication and error correction. In cases were the UMI is short (e.g. 6 bp), basepairs beyond the anchor sequence that contain variability arising from sonication can be added to the UMI. (b) Short reads that cannot be assembled and do not contain an anchor-UMI are processed with the short option of the BabrahamLinkON pipeline. A proxy UMI is constructed from the V end sonication variability and part of the junction in the J end read. The junction sequence by default is taken 50 bp into the J end reads. (c) In cases where short reads do contain an anchor-UMI, this can be used for deduplication much like in the umi option. These reads are processed using the short_anchor option of the pipeline. (d) For long reads that can be assembled, but do not contain an anchor-UMI, the no_anchor option can be used. Much like the short option, the no_anchor option uses the sonication variability and the junctional diversity to construct a proxy UMI.

4.4.1 High homology of J genes results in primer mispriming

Quantitative analysis of clonal counts and clonotype assemblies required a computational solution to several issues not addressed by the previous pipeline. Namely, the mispriming of J sequence results in the wrong annotation of the J genes. The J gene identity is subsequently used for clonotype assembly, which in the case of misprimed sequences leads to the creation of a new clonotype group

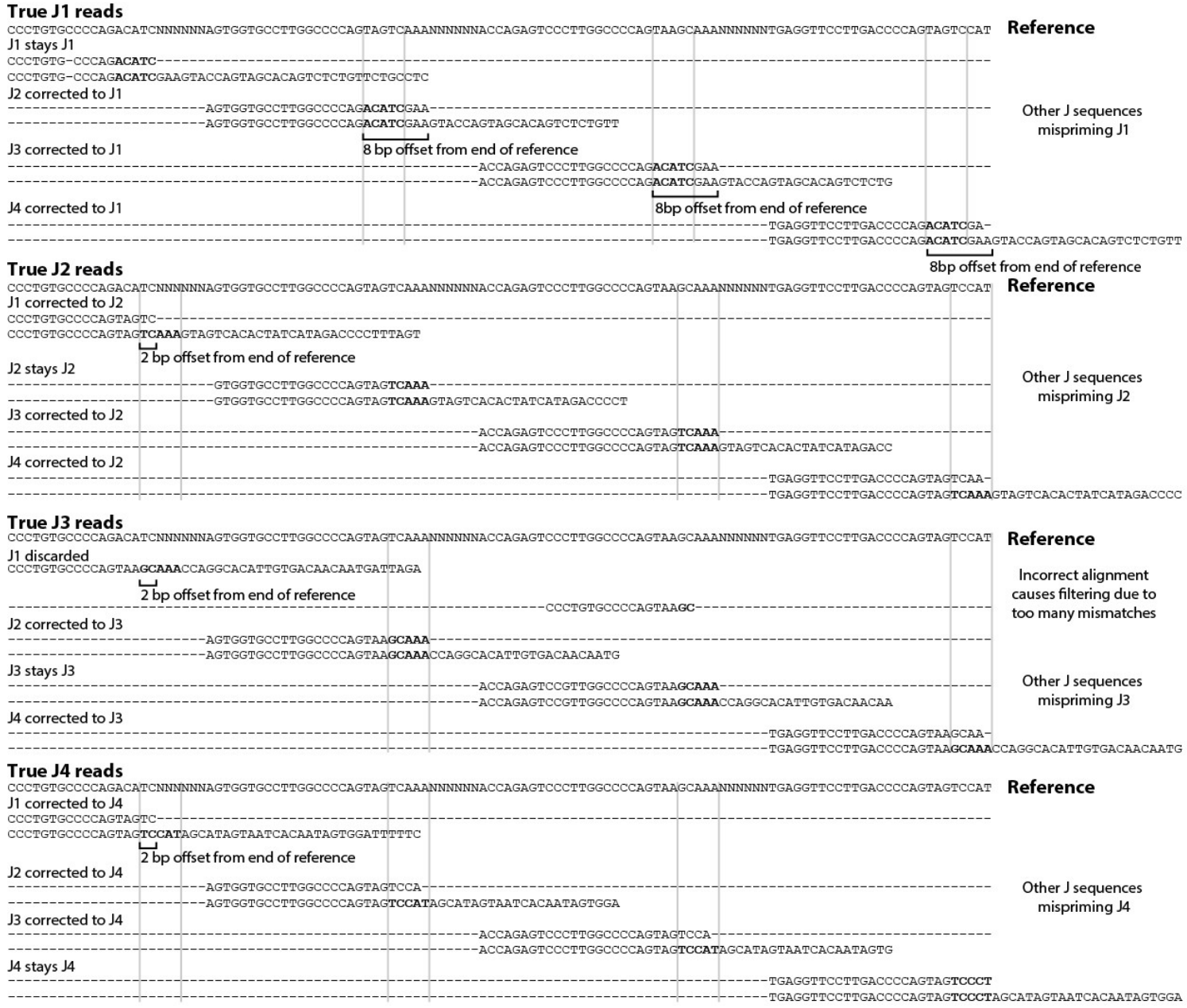
that incorrectly inflates the diversity observed. To better understand the extent of misprimng, I derived the true identity of germline reads by aligning them with Bowtie 2 (Langmead and Salzberg, 2012) to a reference genome. Trimming off the primer sequence is desirable to get a higher quality alignment but is not essential. Subsequently, I aligned the start of each read to a custom reference composed of the J gene primers separated by a string of 6 N characters using the stripped Smith Waterman (SSW) algorithm implemented in the python scikit-bio package (scikit-bio.org) (Figure 4-11). The true identity along with the primer identity allowed me to quantify the extent and characterise the nature of each mispriming event. Most reads fall onto the diagonal, which represent the correct primer amplifying its target J gene (Table 4-1). When I examined the J2 primers, I observed that they were priming the J4 gene (137895) at comparable levels to the target J2 gene (137249). This corresponded to the only mispriming event corrected by the previous pipeline. However, besides the J2-J4 event, other primers were frequently observed within germline sequences of genes they were not intended to amplify. A prominent example of this is the J3 primer amplifying the J4 gene at comparable levels to the J2-J4 mispriming event (Table 4-1). This analysis highlighted the need for a more comprehensive mispriming correction method.

Table 4-1: The extent of mispriming of J genes estimated from germline reads.
The true identity of a germline read is attained from its alignment to one of the four J genes in the mouse. By performing an alignment of the sequence corresponding to the length of the primers, the identity of the primer that amplified a read can be determined. By comparing the true identity with the primer identity, it is possible to quantify the level of mispriming.

Before mispriming correction

		True identity			
		J1	**J2**	**J3**	**J4**
Misprimed	**J1**	253633	3843	84	2030
	J2	1605	137249	50630	137895
	J3	23442	72114	465244	134353
	J4	480	44315	17280	558753

Examining mispriming in germline reads revealed that in most cases the mispriming of each J gene happened in a consistent manner and leaves a portion of the true J gene behind. I chose to use a stretch of 5 bp of the unchanged J gene sequence to find the original J gene identity. The procedure I devised for mispriming correction starts by using the first 5-10 bps of each J end read to identify the J primer (Figure 4-12). The identification is performed using Levenshtein distance (https://github.com/ztane/python-Levenshtein/) which corresponds to the number of operations required to transform one string into another string of equal or different length. The initial identity is used to align the expected length of each J primer to the primer reference. I have found that limiting the length of the sequence being aligned limits false alignments.



(Figure legend on next page)

Figure 4-11: Alignment of the mouse Igh J primer sequences from each read allows the correction of mispriming.
The reference sequence contains along with the J gene primer sequence an additional 5 bp that can be used to identify the misprimed J gene. Each J primer sequence is separated by a string of 6 N characters. By aligning germline reads to the reference genome it is possible to determine the original J gene that was extended and captured. The alignment of the primer sequence within each read was performed using the stripped Smith Waterman algorithm. This allows the identification of sequences that contain insertions/deletions (indels) or mismatches as illustrated by the J1 primer in the J1 true reads. The sequence identifying the true J identity is highlighted in bold with vertical grey lines tracing it back to the reference.

If the initial J identity matches the aligned reference J, a set of offsets are used to locate the position of the 5 bps that match the true J gene. The offsets are determined by the distance from the end of each J reference to the start of the 5 bp used for mispriming correction (Figure 4-11 and Table 4-2). For every sequence with an identifiable J primer, a 5 bp sequence is extracted for each J gene based on the expected location of the true J gene sequence, which is subsequently compared with a 5 bp reference for that J gene. The comparison is performed using Levenshtein distance and the gene with the 5 bp sequence of zero distance is used to correct the primer sequence. Primer sequences with indels and mismatches are also corrected (Figure 4-11) and the final identity of the J sequence is written into the sequence read name (Figure 4-8).

During mispriming correction there is a small number of reads that escape correction and are incorrectly assigned a J gene sequence (Table 4-3). In rare cases where two or more J genes have a perfect match 5 bp sequence, the reads are marked as multi-hit sequences (Figure 4-12) and are filtered out by default at the beginning of deduplication. In addition, a certain proportion of sequences that do not have a perfect match to any of the 5bp J gene reference sequences are marked as unclear (Figure 4-12). The lack of a perfect match could be caused by several mechanisms such as PCR and sequencing error within the 5 bp sequence or chewback from the NHEJ machinery (discussed in the next section). Because these unclear sequences still represent legitimate recombination products they are by default retained in downstream analysis. The J gene mispriming takes place during the first round PCR (Figure 4-1), which means a single UMI group will contain a mixture of both correct and misprimed sequences. Because the misprimed sequences are typically less frequent than correctly primed sequences, especially after mispriming correction, it is expected that by collapsing a group of sequences within a UMI group into a consensus sequence the J misprimed sequence minority will be corrected. Indeed, the analysis of the true vs misprimed reads after deduplication and consensus sequence construction shows that only a small number of incorrect J identity reads remain (Table 4-3). The same is true for the germline unclear reads that contain mismatches in their 5 bp mispriming correction sequence (Table 4-4). Filtering UMI groups with low number of reads at the deduplication stage can also increase the rate at which the consensus sequence returns the correct J primer, but the ability to do this will depend on sequencing depth.

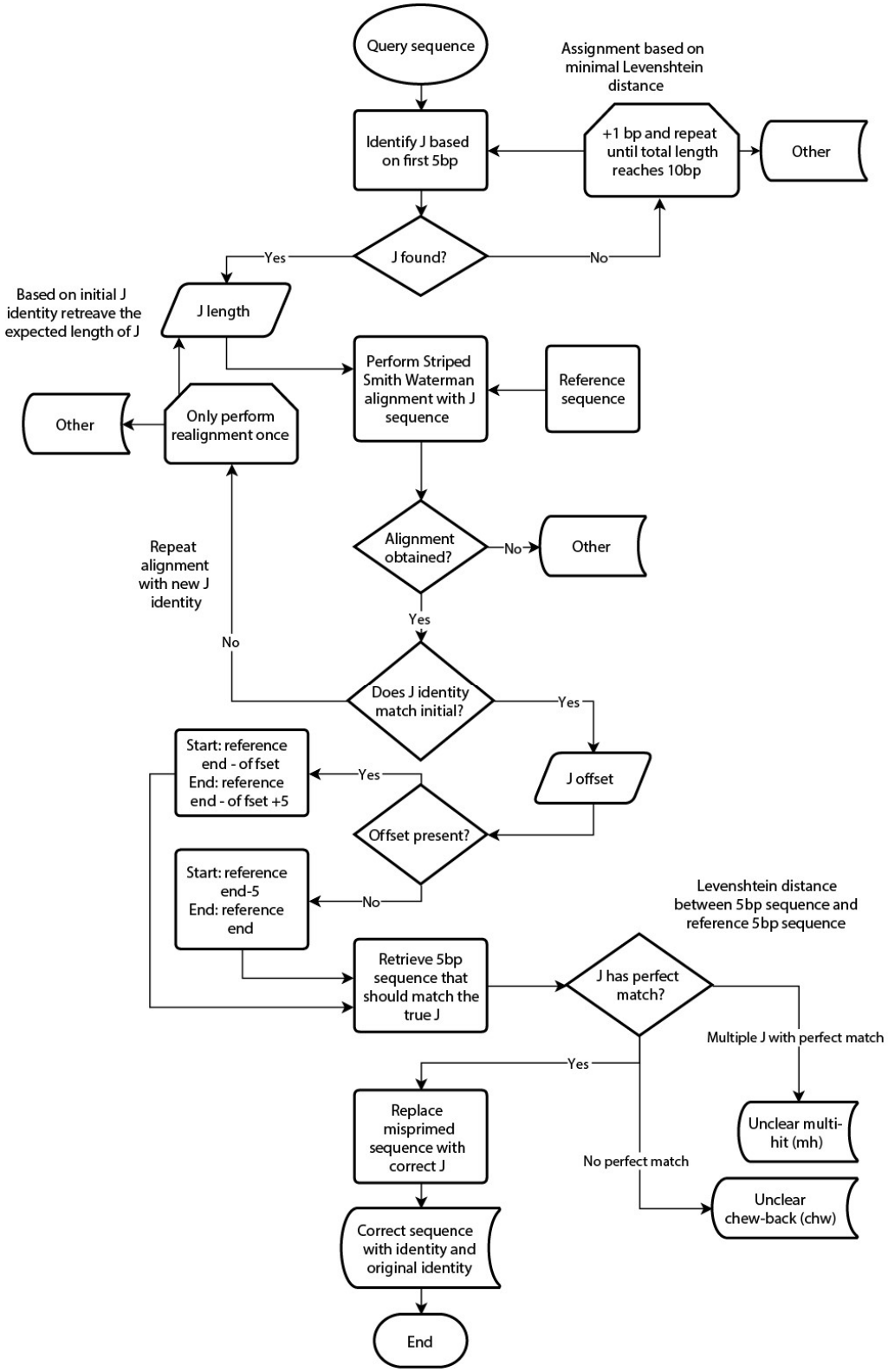


Figure 4-12: Flowdiagram describing the process of mispriming correction.
The correction of misprimed sequences is based on a 5bp sequence beyond the primer that can be unambiguously identify the true J gene. Based on how the sequence is misprimed, the 5bp is in different, but consistently reproducible positions along the read. By using a set of offsets, it is possible to find the location of the 5bp along the misprimed read. A reference sequence is created from the primer sequences (+5bp beyond primer), with each J sequence separated by 6 N bases from each other. The alignment to a reference sequence is performed using the striped Smith Waterman algorithm from the Scikit-bio package (http://scikit-bio.org). After the identification of the true J, the primer sequence is replaced with the correct J sequence.

Table 4-2: A set of offsets for each immunoglobulin locus is used for mispriming correction.
For the mouse Igh locus the set of offsets is primarily linked to the J1 gene, which misprimed other J gene with a 2 bp offset while getting misprimed itself with an 8 bp offset (see Figure 4-11 for graphical explanation). For the mouse Igκ locus an additional primer identity sequence had to be introduced due to the J4 primer mispriming J5 with an alternative 5 bp identification sequence (see Figure 4-13 for a graphical explanation). The human IGH locus is by far the most complicated to correct due to the higher number of J genes and their multiple alleles. Numerous alternative 5 bp identification sequences had to be introduced and for J6 it was not possible to distinguish between the different alleles.

Mouse Igh mispriming offsets

		Misprimed J offset			
		J1	**J2**	**J3**	**J4**
Primer identity	**J1**	-	2	2	2
	J2	8	-	-	-
	J3	8	-	-	-
	J4	8	-	-	-

Mouse Igκ mispriming offsets

		Misprimed J offset			
		J1	**J2**	**J4**	**J5**
Primer identity	**J1**	10	-	-2	9
	J2	-	-	-	-
	J4	-	-	-	-
	J5	-	1	3	-
	J5.c	-	-	-	-

Human IGH mispriming offsets

		Misprimed J offset							
		J1	**J2**	**J3**	**J4.1**	**J4.3**	**J5.1**	**J5.2**	**J6.1**
Primer identity	**J1**	-	4	6	1	-	7	7	-
	J2	-	4	6	1	-	7	7	-3
	J3	4	3	-	0	-	6	6	-4
	J4.1	0	-1, -2	1	-	-	2	2	-8
	J4.3	0	-1, -2	1	-	-	2	2	-8
	J4.1c	1	0, -1	1	1	-	3	3	-7
	J4.3c	1	0, -1	1		-	3	3	-7
	J5.1	3	2	4	8	7	-	-	-5
	J5.2	3	2	4	8	7	-	-	-5
	J6.c	0	-	-	-	-	-	-	0
	J6.c2	0	-	-	-	-	-	-	0
	J6.3	-	-	-	-	-	-	-	-
	J6.4	-	-	-	-	-	-	-	-

Table 4-3: Success of mispriming correction judged from germline reads.
The success of the correction can be seen by running the misprimed corrected reads through the germline alignment and primer identification script again. The few reads present off the diagonal show that the correction successfully replaced the J primer sequence with the true J sequence in most cases. Because mispriming takes place during the first PCR amplification, a sequence marked with a single UMI can have multiple different J primers mispriming it. Therefore, since most sequences are primed by the correct J primer, by collapsing a group of sequences, based on their UMI, into a consensus sequence it can be expected that this will correct the misprimed J minority of reads. Indeed, by analysing the true vs misprimed identity of reads after deduplication and consensus sequence construction it can be seen that only a small number of reads remain off the diagonal.

After mispriming correction

		True identity			
		J1	**J2**	**J3**	**J4**
Misprimed	**J1**	279129	3	5	10
	J2	8	257441	435	69
	J3	10	62	532775	17
	J4	13	15	23	832935

After deduplication and consensus sequence

		True identity			
		J1	**J2**	**J3**	**J4**
Misprimed	**J1**	4100	0	0	0
	J2	0	5878	10	0
	J3	0	0	8953	1
	J4	2	1	4	16478

Table 4-4: Unclear mispriming events tend to have correct identity after deduplication.
The lack of the 5 bps used for mispriming correction in some sequences results in the inability to confidently resolve the identity of these events, which are marked as unclear. However, the correct identity is still obtained in most cases after the generation of a consensus sequence.

Unclear before deduplication

		True identity			
		J1	**J2**	**J3**	**J4**
Misprimed	**J1**	9989	32	52	25
	J2	19	1571	331	1613
	J3	254	424	7324	1503
	J4	7	276	127	15908

Unclear after deduplication and consensus sequence

		True identity			
		J1	**J2**	**J3**	**J4**
Misprimed	**J1**	1	0	0	0
	J2	0	99	1	5
	J3	0	8	229	3
	J4	0	0	0	37

For the mouse and human Igh loci clear majority of mispriming events can be successfully corrected. The exception being the J1 primer mispriming the J3 genes in the mouse Igh, where the short J1 primer

sequence is incorrectly aligned to the J3 reference and is discarded due to too many mismatches (Figure 4-11). The Igκ poses additional challenges, as several mispriming events cannot be corrected and result in false J gene identities. A prime example is the J1 gene being misprimed by J2-5 primers (Figure 4-13). The CCTCC sequence used to identify the J1 gene is overwritten by a stretch of C nucleotides present within the J2-5 primers, which in the case of J2 and J5 primers produces a sequence that is indistinguishable from a true J2 and J5 gene. In addition, there are several examples of J genes being primed in a way that somehow results in two or more sequence variants at the expected location of the 5 bp used for mispriming correction (Figure 4-13). In the case of the J4 primer mispriming the J5 gene, an additional reference sequence (J5c) was introduced for mispriming correction based on a large fraction of reads displaying this variant. Because of the inability to correct certain mispriming events, I introduced a metric based on the germline reads called the mispriming error (Table 4-5, Table 4-6, Table 4-7). The metric allows users to examine what the error of mispriming correction is for different species and loci and allows users to make informed decision regarding downstream analysis. The error for the mouse and human IGH is minimal (Table 4-5 and Table 4-7), while for the mouse Igκ the inability to correct certain mispriming events leads to a higher percentage error (Table 4-6). With increasing J gene and allele numbers the mispriming correction becomes even more complicated, as exemplified by the high number of additional reference sequences required for correction of human IGH sequences (Table 4-2). Because of the repetitiveness of the immunoglobulin loci, it is perhaps not surprising that the 5 bp sequence used for mispriming correction can be observed at multiple locations. The repetitiveness underscores the importance of looking in the correct location for the 5 bp sequence and in rare cases it can also lead to misidentification. A good example is the 5 bp sequence beyond the J1 primer in the Igκ locus (ACCGA), which also appears in the J5 gene after a J1 primer has misprimed it (Figure 4-13). As a result, a sequence within the J1 primer had to be used (CCTCC), which prevented the correction of a proportion J1 gene sequence misprimed by the J2 primer. In cases where the J identity percent error after mispriming correction is high, performing clone assembly into clonotype using only the CDR3 sequence and the V gene identity may be desired.

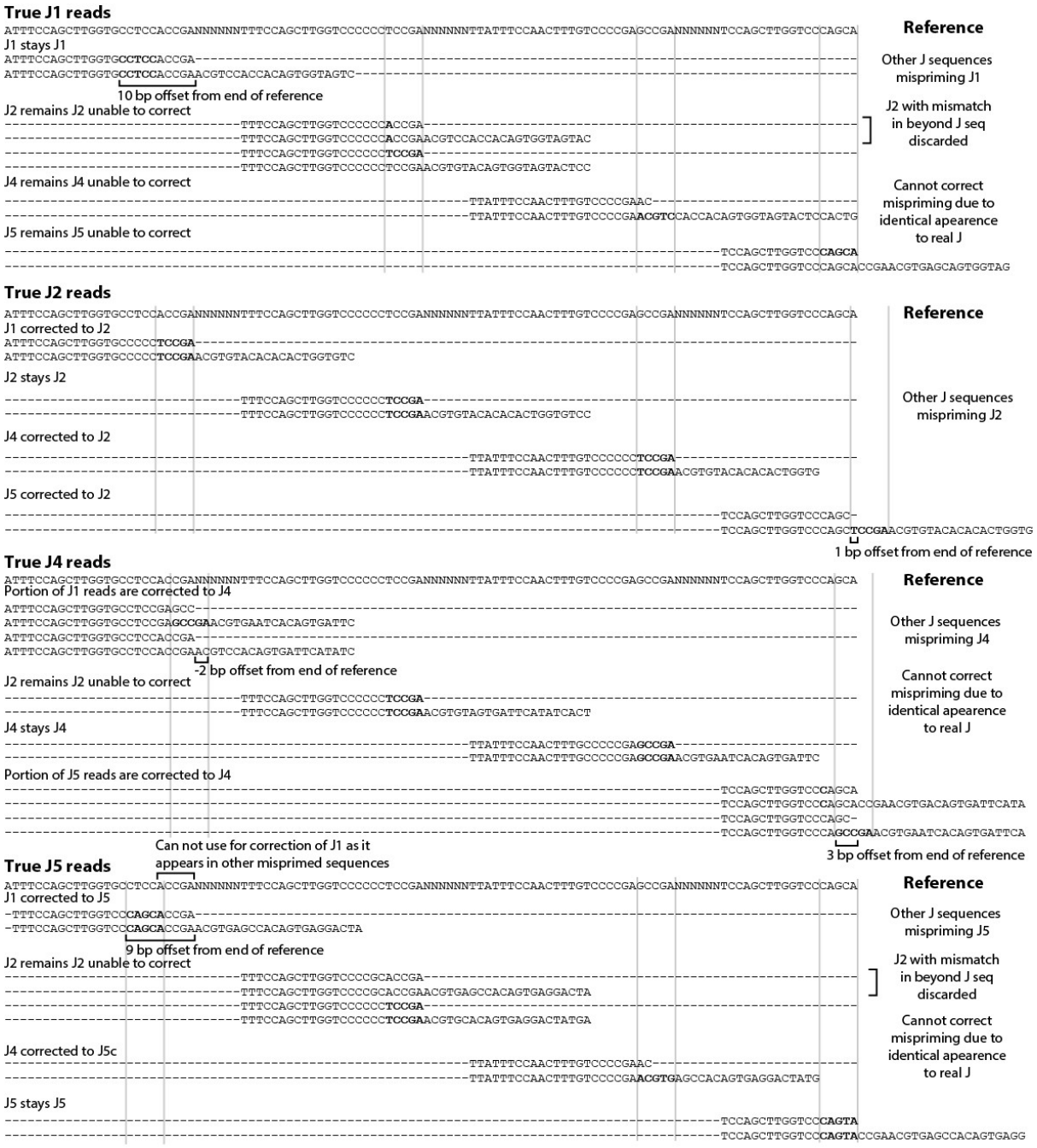


Figure 4-13: Alignment of the mouse Igκ J primer sequences reveals the inability to correct certain mispriming events. In the case of J1 gene mispriming by J2-5 primers, the CCTCC sequence used to identify the J1 read is not present as the single T is replaced by the stretch of Cs within the J2-5 primers. Unlike other J genes, the 5 bps beyond the J1 primer sequence cannot be used for mispriming correction as this sequence also appears in the J5 gene that is misprimed by J1. Another issue arises for the J4 and J5 genes, where mispriming with the J2 primer looks identical to a true J2 read. The mouse Igκ locus also contains examples were the 5 bp sequence is not conserved for all reads, as illustrated by J2 mispriming of J1 reads, J1 and J5 mispriming of J4 reads and J2 mispriming of J5 reads. Because the 5 bp does not match any J genes, these sequences will be labelled as unclear.

Table 4-5: The extent of J mispriming before and after mispriming correction for the mouse Igh.
Percent error is the portion of incorrectly identified J genes after mispriming correction, based on germline sequence, over the total J gene reads. The mispriming error may vary from batch to batch, so it can useful to examine the mispriming error from the germline reads of each sample just as a quality control.

Igh before mispriming error			
	Total	Misprimed	Percent misprimed
J1	4074	27	0.662739
J2	6379	470	7.367926
J3	10419	1465	14.06085
J4	14587	1890	12.95674

Igh after mispriming error			
	Total	Misprimed	Percent error
J1	4101	0	0
J2	5919	10	0.168947
J3	8955	1	0.011167
J4	16484	7	0.042465

Table 4-6: The extent of J mispriming before and after mispriming correction for the mouse Igκ.
Percent error is the portion of incorrectly identified J genes after mispriming correction, based on germline sequence, over the total J gene reads.

Igκ before mispriming correction			
	Total	Misprimed	Percent misprimed
J1	42435	11020	25.97
J2	58363	1639	2.81
J4	74006	1167	1.58
J5	44053	4069	8.89

Igκ after mispriming correction			
	Total	Misprimed	Percent error
J1	32935	1551	4.71
J2	61868	1703	2.75
J4	74870	211	0.28
J5	50750	506	1.00

Table 4-7: The extent of J mispriming before and after mispriming correction for the human IGH.
Percent error is the portion of incorrectly identified J genes after mispriming correction, based on germline sequence, over the total J gene reads.

Human IGH before mispriming correction			
	Total	Misprimed	Percent misprimed
J1	54645	47861	87.59
J2	24101	13929	57.79
J3	103284	87109	84.34
J4	54324	32679	60.16
J5	177456	40728	22.95
J6	219413	24	0.01

Human IGH after mispriming correction			
	Total	Misprimed	Percent error
J1	11781	113	0.96
J2	16984	19	0.11
J3	30513	50	0.16
J4	47846	116	0.24
J5	331893	51	0.02
J6	313242	45	0.01

4.4.1.1 Endonuclease chew back in NHEJ DNA repairs can extend into the 5bp used for mispriming correction

The non-homologous end joining (NHEJ) repair machinery poses an additional challenge to correcting J gene mispriming events. The nuclease activity of NHEJ, which is designed to expose short stretches of microhomology that can facilitate end joining, results in the nibbling (chew-back) of the J gene sequence (Chang et al., 2017; Malu et al., 2012). To determine the extent of chew-back within each J gene, I created sequence logos that display the information content at each position after the 5 bps that are used for mispriming correction. When the same base is always present at an observed position, the information content of this position would be the maximum value of 2 bits (this is because a single base at a position for DNA tells us two bits of information about itself, that is whether it is a pyrimidine or a purine and whether it is either A/G or C/T). If two bases are observed at equal frequency at a position than the information content of this position would be a total of 1 bit. In other words, the height of the bases can be interpreted as the degree of conservation of that residue at a position within a sequence. In the case of the 5 bps, which are required to be present in every sequence that can be misprimed corrected, the information content is 2 bits (Figure 4-14). With each subsequent position the information content decreases as the conserved germline sequence is replaced by other bases due to the nuclease activity of NHEJ. This is nicely observed for the J1 and J4 sequences, where a gradual decrease in information content is observed when moving away from the

first 5 bases (Figure 4-14). The effect of chew-back is most prominent at the RSS heptamer where the double-stranded break is created (Figure 1-3 a). In the case of J2 and J3 the much sharper transition between the 5 bp mispriming correction sequence and the chew-back sequence is due the proximity of the cleavage site. This means that the 5 bps used for mispriming correction of J2 and J3 will be conserved less frequently than in J1 and J4.

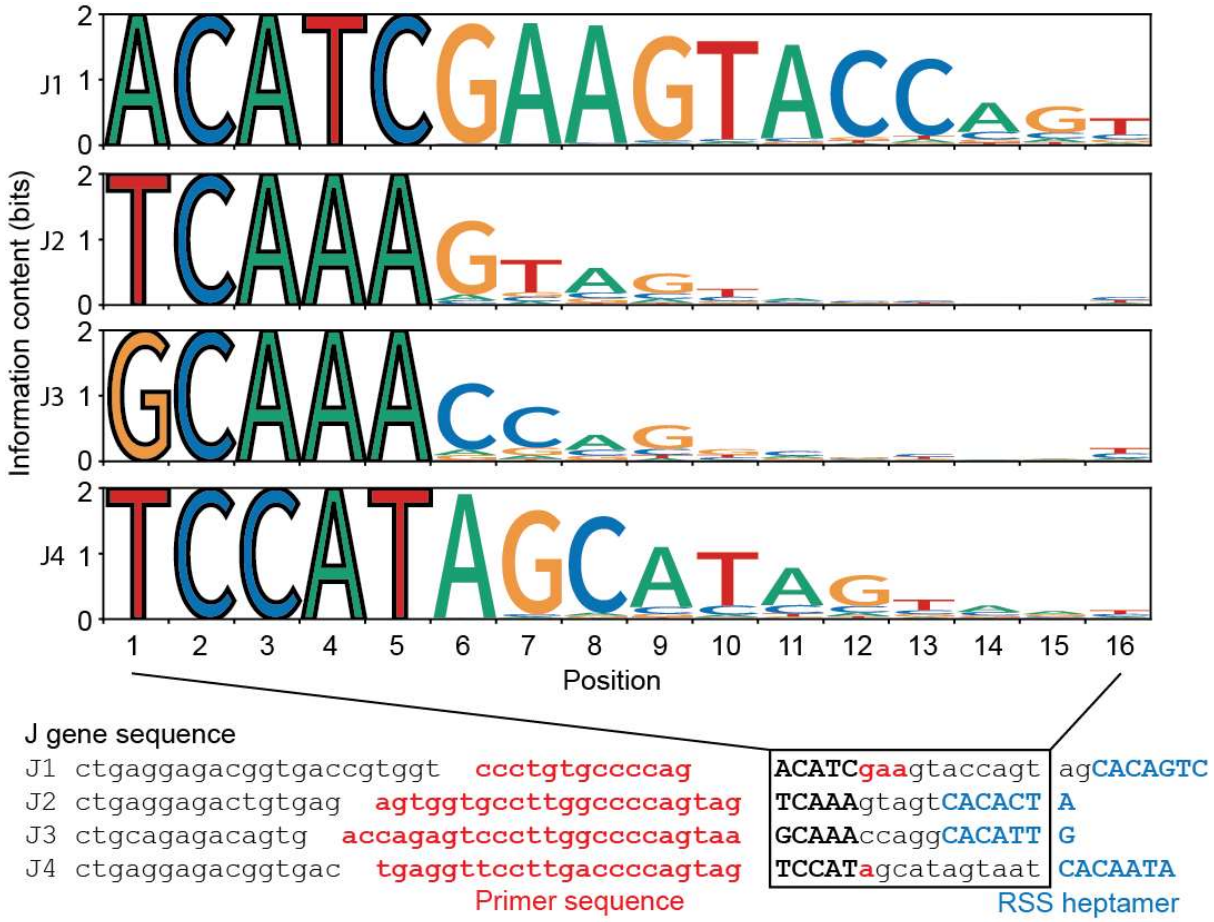


Figure 4-14: Seqlogo illustrates the impact of chew-back around the sequence used for J gene mispriming correction. The first five positions of the seqlogo correspond to the basepairs used to identify the true J gene and correct mispriming sequences. Below the seqlogo are the four mouse J gene sequences with the primers highlighted in red. The high degree of homology can be seen within the primer sequences. The sequence within the rectangle matches the position of the seqlogo, with the bold capital letters corresponding to the 5 basepairs used for mispriming correction. It is possible to use these 5 basepairs because they are either located beyond the primer sequence, as in the case of J2 and J3, or remain after other J primer have misprimed, as in the case of J1 and J4. The nuclease chew-back from the non-homologous end joining machinery is most prominent at the RAG1/2 cleavage site and diminishes with increasing distance. The cleavage site is located at the start of the RSS heptamer sequence highlighted in blue. The seqlogos were created using the matplotlib python library (Hunter, 2007) with the approach described by Saket Choudhary and Markus Piotrowski (https://github.com/saketkc/motif-logos-matplotlib/blob/master/Motif%20Logos%20using%20matplotlib.ipynb).

4.4.2 Correcting errors within UMI sequences

The presence of PCR and sequencing errors within UMI sequences can unintentionally lead to increased repertoire diversity. For this reason, I decided to explore methods for UMI correction and implement one into the deduplication pipeline. The presence of UMI errors has been previously documented and removing UMI groups with reads counts below a certain threshold was proposed as

a possible solution (Islam et al., 2014). A study conducted with iCLIP and single cell RNA-seq datasets compared five different method for the correction of errors within UMI sequences, including the read count threshold method (Smith et al., 2017). Three of the methods in the study were developed by the authors, who demonstrated that their directional-adjacency method outperformed all the other methods. The premise behind directional-adjacency is that errors within the UMI sequences arise during the library making process from correct UMI sequences that are present at much higher frequency. This bridges the idea of removing low count UMI groups that correlate with errors within the UMI sequence and the low divergence of related UMIs. By connecting UMI groups (nodes) that are different by a single nucleotide (adjacent nodes) and can be grouped into a structure where the head node has a read count that is more than 2n-1 of the adjacent node (directionality of the edges), a network of connected UMI sequence is created (Figure 4-15). Each network contains a head node that represents the highest count UMI group under which the UMIs with errors can be collapsed. This approach of sequence error correction has also been used in the immunoglobulin annotation tools MiXCR for correcting errors within CDR3 sequence and collapsing clonotypes (Bolotin et al., 2015). As a result, I chose to implement the directional-adjacency UMI error correction method into BabrahamLinkON.

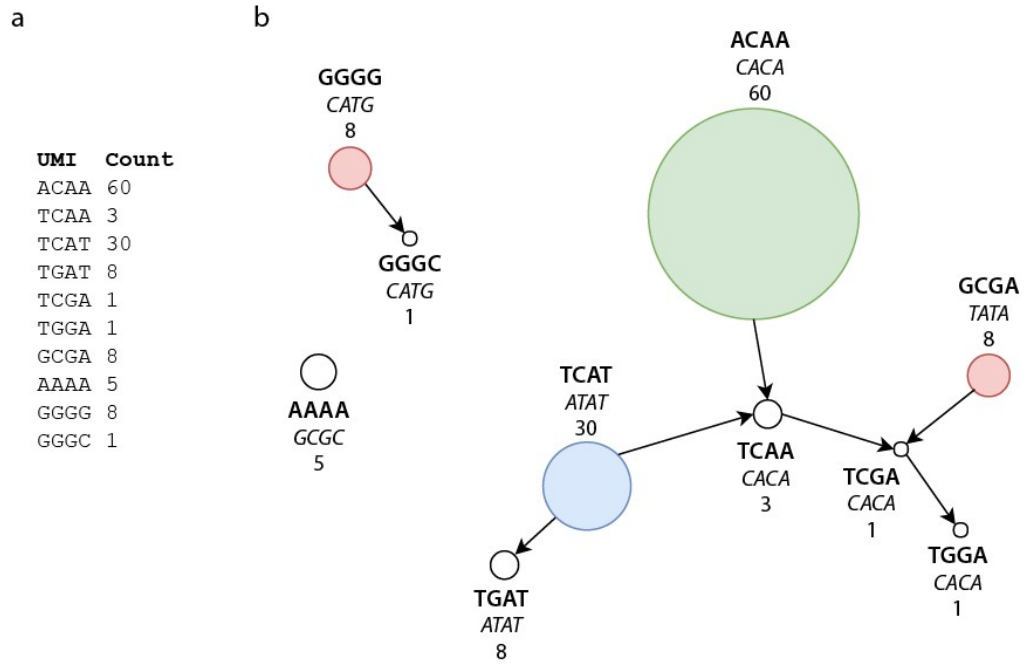


Figure 4-15: Constructing networks of UMI sequences for UMI error correction based on directional adjacency.
(a) The network is constructed from a list of UMI and their counts using two rules. A UMI can only have 1 mismatch from an adjacent UMI and the head node UMI (one with the highest count) must have a count of more than 2n-1 compared to the adjacent UMI. This strategy for UMI error correction was devised by Tom Smith et. al. in UMI-tools (Smith et al., 2017). (b) Each node represents a UMI group with the UMI sequence shown in bold. In italic is the read sequence, which would be much longer in reality, with the number of reads within each UMI group displayed underneath. The size of the nodes is proportional to the UMI group sequence count. The directional networks all contain a head node (highlighted in red, blue or green) with adjacently connected components. The idea is to collapse UMI groups with lower read count under a higher count head node from which the UMI error arose. However, in some cases networks with multiple head nodes arise that need to be resolved.

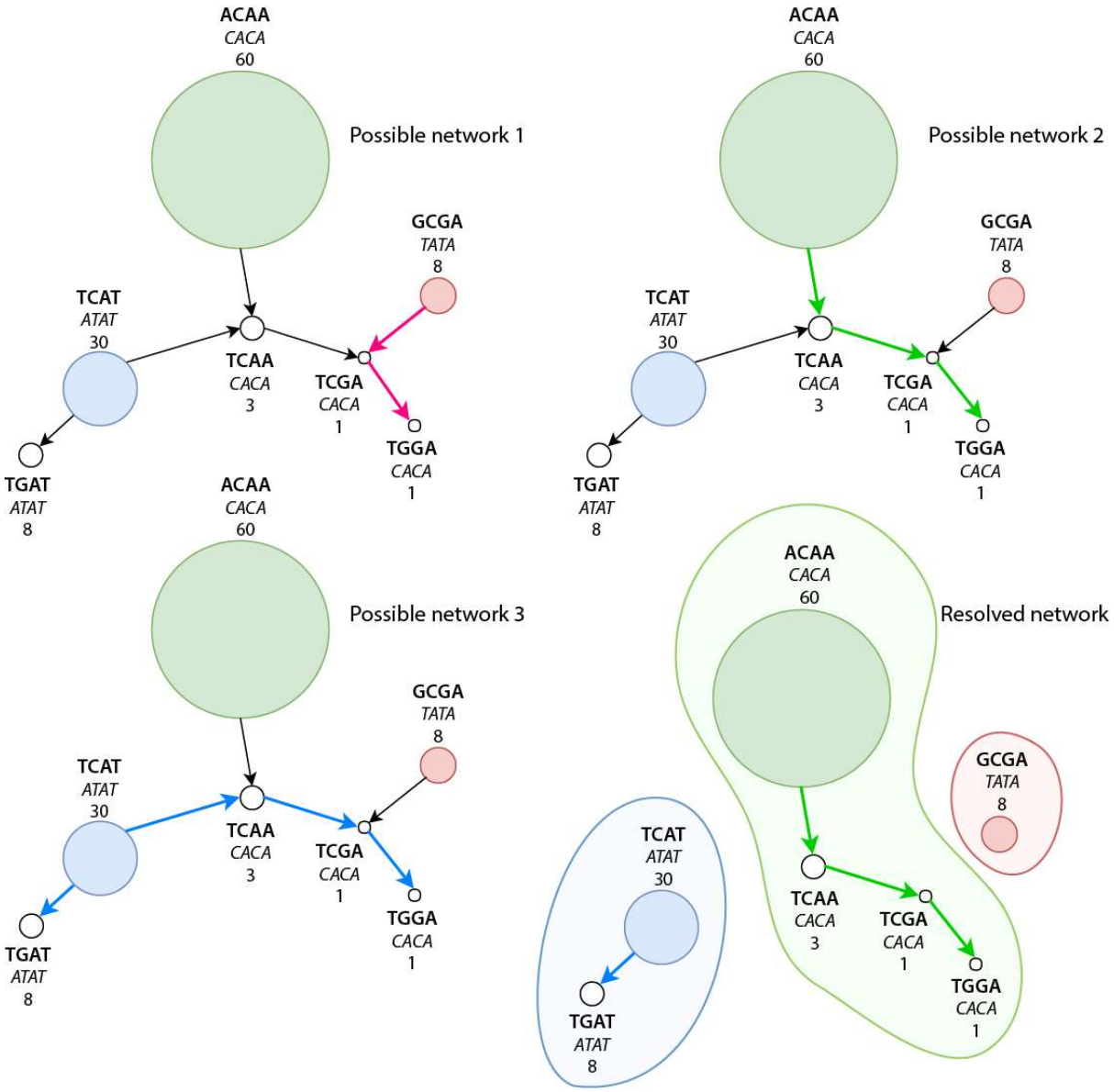


Figure 4-16: Resolving networks with multiple head nodes is done using the consensus sequence of each UMI group. In this example there are three possible networks formed from each of the head nodes (red, green and blue). The connected component within each network are based on the directionality of the edges (2n-1) and adjacency of node (1 mismatch between UMIs) (each possible network is highlighted by coloured edges that match the head node colour of that network). Several adjacent nodes are present within multiple networks and need to be assigned to a singular head node. To resolve to which head node the adjacent nodes belong to, the consensus sequence of each adjacent node overlapping multiple networks is compared to the consensus sequence of each head node. The head node that best matches the adjacent node consensus sequence is assigned that adjacent node. At the end, a separate network containing only a single head node is obtained, under which all the adjacent node sequences can be collapsed.

While implementing the directional-adjacency method from UMI-tools for VDJ-seq UMIs, I ran into the problem of a single network containing multiple head nodes (Figure 4-15). I decided to resolve the networks by comparing the consensus sequence of the head nodes with the consensus sequence of the shared adjacent node and assigning the adjacent nodes to the best head node match. In RNA-seq and iCLIP datasets the UMI groups are additionally split depending on their alignment across the

genome. Only UMIs of reads that align to the same region (gene) are considered for correction, meaning the number of encountered UMI is drastically reduced (Smith et al., 2017). This suggested to me that perhaps because in VDJ-seq the UMIs are only split into the two anchor groups, the diversity of the UMIs may be on the limit. I examined the UMI diversity within a 5 µg starting material dataset by placing each UMI onto the full diversity space created by a 12N UMI (over 16 million different combinations) (Figure 4-17). The analysis revealed that indeed there were biases in composition of observed UMIs, which I interpreted to be the result of the manufacturing process. By using a mixture of UMIs manufactured separately (CV 0.37 and CV 0.43), afforded by the two anchors, the random bias of the two batches was reduced (CV 0.25; Figure 4-17). This highlighted yet another unexpected role of the two separate anchor sequences. Overall, the observed spikes in very closely related sequences can explain the prevalence of multi-head node networks.

The splitting of UMIs by only the anchor sequence, instead of by genomic locations like in RNA-seq, raised another issue with large datasets. Due to the non-linear increase in computational time with increasing UMI numbers, the UMI error correction of large samples takes a very long time despite my attempts at optimising several key algorithms. Even though this makes it impractical to routinely use UMI error correction, the procedure still holds value in evaluating the length of proxy UMI in downsampled datasets. Because the diversity of a proxy UMI is unknown, it is difficult to decide how long it should be. By testing a range of V end and J end UMI lengths using a 200,000 read downsampled file, with UMI correction either on or off, I was able to empirically determine an optimal length of the proxy UMI (Table 4-8). With short proxy UMI lengths (8-6 and 8-7) the number of output reads from the UMI correction is consistently lower compared to the non-corrected reads. This suggests that the UMI correction is correcting UMIs that should not be corrected, leading to their subsequent filtering with the gt-ratio filter. This points to a low diversity proxy UMI. Alternatively, with long proxy UMIs (8-9) the number of output sequences is now higher than the output without correction (Table 4-8). The longer the UMI, the higher the probability that it will contain a PCR or sequencing error. The higher output reads from the UMI correction suggest that in these long proxy UMIs there are now errors that are being successfully corrected. However, as the intention is not to use UMI correction for the full dataset the long proxy UMI are not ideal. The intermediate UMI length (8-8) shows a near identical read output from both corrected and not corrected UMIs, suggesting a balance between UMI diversity and UMI errors has been achieved. In addition to UMI length, I also tested two cut-offs for low UMI group read count. The cut-off of 1, which only removed UMI group with a single sequence, was insufficient as with increasing proxy UMI length the impact of PCR and sequencing errors was masked by the low count groups that are most likely errors themselves.

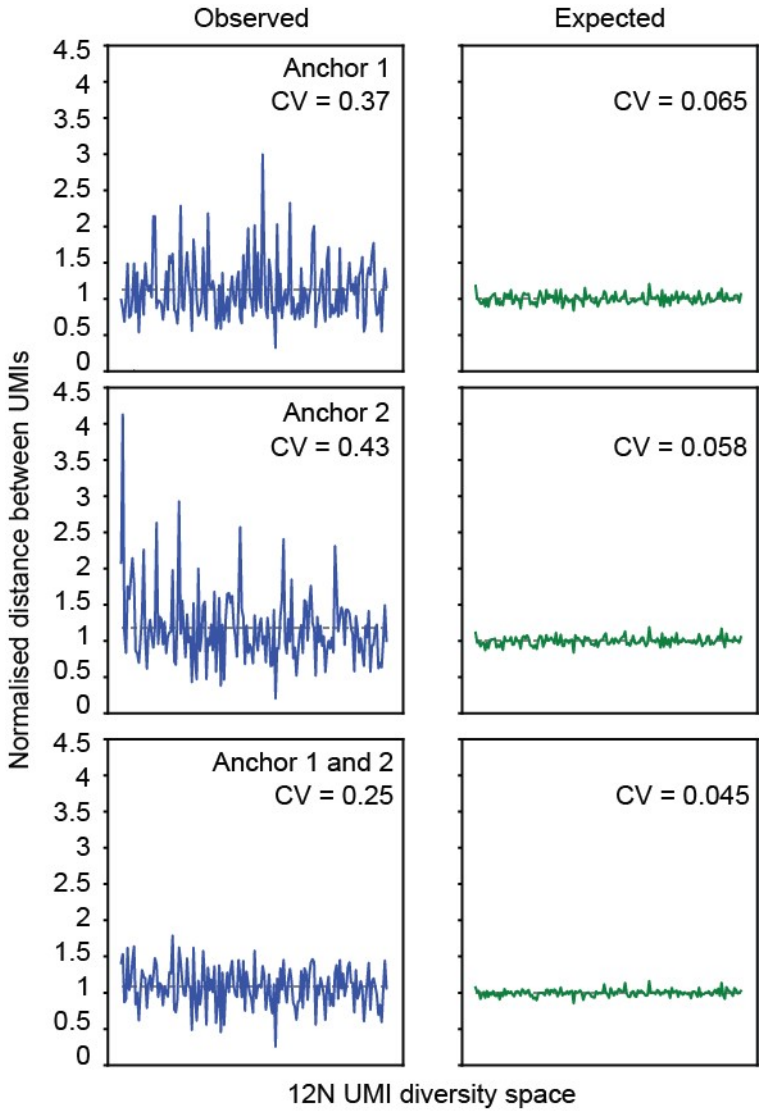


Figure 4-17: Distribution of UMI sequences displays biases towards certain type of composition.
The plots represent the entire diversity space for a 12N UMI sequence (4^{12} – over 16 million possibilities) divided into 166 individual bins. Each bin is the mean distance between sampled UMIs. The distance is calculated by measuring the number of unsampled UMIs between two sampled UMIs. Each plot contains a different number of UMIs depending on the content within the 5 µg dataset used for the observed plots (blue). To maintain an equal scale between the plots, each plot was normalised by the distance between equally distributed UMIs. The expected distance between UMIs was calculated by randomly sampling an equal amount of UMIs as in the observed sample from the entire diversity space (expected green plots). The distance between perfectly distributed UMIs if shows as a grey dashed line, which also represents the mean of all the bins. The degree of the bias is shown by the distance away from the dashed line (the further away the greater the bias). The bias seen within the UMI distribution varied between the two anchors suggesting it varies for each manufactured oligo. The spikes above the dashed line highlight areas were very few UMI are sampled from, leading to an underrepresentation of these UMI within the sample. On the other hand, the spikes below the dashed line highlight areas of highly sampled UMIs, which leads to an overrepresentation with the sample. A coefficient of variation calculated for each plot shows that the joint use of both anchor sequences together reduces the biased representation of UMIs within the final sample.

Table 4-8: UMI correction procedure reveals the optimal length of proxy UMIs.
For sequences without a UMI or a very short UMI the use of the sonication variability (V end length), along with the junctional diversity (J end length) can be used together as a proxy UMI. However, the difficulty with using a proxy UMI is in the choice of what length to use. A proxy UMI that is too short will not provide enough diversity and group incorrect sequences, while a proxy UMI that is too long will suffer from the increase chance of encountering PCR and sequencing errors, which will artificially inflate diversity. Values that are less important have been greyed out.

Cut 1	Uncorrected	UMI corrected	Uncorrected	UMI corrected	Uncorrected	UMI corrected	Uncorrected	UMI corrected
V end length	8	8	8	8	8	8	8	8
J end length	9	9	8	8	7	7	6	6
Output reads	9625	9250	9576	9086	9501	8680	9404	8203
Cluster discarded	1566	2746	1755	3125	1979	3802	2247	4565
Low UMI count	44197	41109	44026	40532	43764	39451	43477	38042
Directional-adjacency corrected	0	1947	0	2184	0	2688	0	3326
Corrected clusters with low ratio	0	1422	0	1682	0	2259	0	2943

Cut 2	Uncorrected	UMI corrected	Uncorrected	UMI corrected	Uncorrected	UMI corrected	Uncorrected	UMI corrected
V end length	8	8	8	8	8	8	8	8
J end length	9	9	8	8	7	7	6	6
Output reads	2392	2416	2390	2394	2393	2236	2383	2067
Cluster discarded	1566	2746	1755	3125	1979	3802	2247	4565
Low UMI count	51430	47943	51212	47224	50872	45895	50498	44178
Directional-adjacency corrected	0	1947	0	2184	0	2688	0	3326
Corrected clusters with low ratio	0	1422	0	1682	0	2259	0	2943

Table 4-9: Number of output reads after deduplication of VDJ-seq libraries prepared with different number of primer extension cycles.
With increasing number of primer extension cycles the number of output reads counterintuitively goes down. However, the number of low count UMI groups that are filtered out increases, suggesting that libraries with higher number of cycles were insufficiently sequenced.

	2 cycles 1	2 cycles 2	4 cycles 1	4 cycles 2	8 cycles 1	8 cycles 2	10 cycles 1	10 cycles 2	12 cycles 1	12 cycles 2
Input reads	431950	556415	474362	581471	777151	841914	842689	1047160	878002	1196872
Output reads	35061	41403	43131	53028	26082	29485	7704	12312	1136	2464
Low ratio discarded	10304	12668	13841	15654	26079	29489	31507	41916	35890	56917
Low count groups	21773	25483	80731	89729	384375	405991	560108	659556	685058	886440

4.4.3 Increasing the number of primer extension cycles results in template switching

During VDJ-seq optimisation we wanted try and increase the capture efficiency that is only around 7-10%, depending on starting material. Because VDJ-seq was always performed with only a single cycle of primer extension, Dan Bolland tested the impact of increasing numbers of cycles on the final library diversity. The initial logs from the deduplication portion of the BabrahamLinkON pipeline revealed decreasing number of output reads with increasing numbers of cycles (Table 4-9). This counterintuitive result made sense considering the increasing numbers of filtered low count UMI groups with increasing cycles. Due to the higher diversity of reads within the higher cycle samples, the current sequencing depth was most likely insufficient despite devoting a larger portion of the sequencing lane to higher cycle samples (Table 4-9 Input reads). As a result, I decided to include the low count UMI groups in further downstream analysis. The increased primer extension cycles produce more UMI unique reads, but surprisingly after performing an additional deduplication using the entire length of the read, many sequences were removed (Figure 4-18). By examining one of the series of duplicated sequences I noticed that some contained two copies of the anchor sequence which should not be able to happen based on the adaptor design (Figure 4-19). The large number of duplicated sequences with different UMIs also prompted me to examine the number of clones within each clonotype. Given that all the sample are technical replicates from the same mouse naïve spleen (unstimulated by antigen), the size of clonotype groups is expected to be low and equal between all samples. Instead, an increase in the size of clonotype groups correlating with increasing cycle numbers is observed (Figure 4-20). Altogether, the double anchors and the unique UMIs in sequences that are expected to have the same UMI, points to template switching. This is additionally supported by the increasing number of sequences being filtered out by the gt-ratio filter (Table 4-8 Cluster discarded). Template switching has been previously documented as the cause of chimeric reads and could explain replacement of UMIs by those originating from free adaptors or other DNA molecules which could be exacerbated by the repetitive nature of the immunoglobulin loci (Patel et al., 1996). For this reason, it is recommended to only perform a single primer extension cycle (Chovanec et al., 2018).

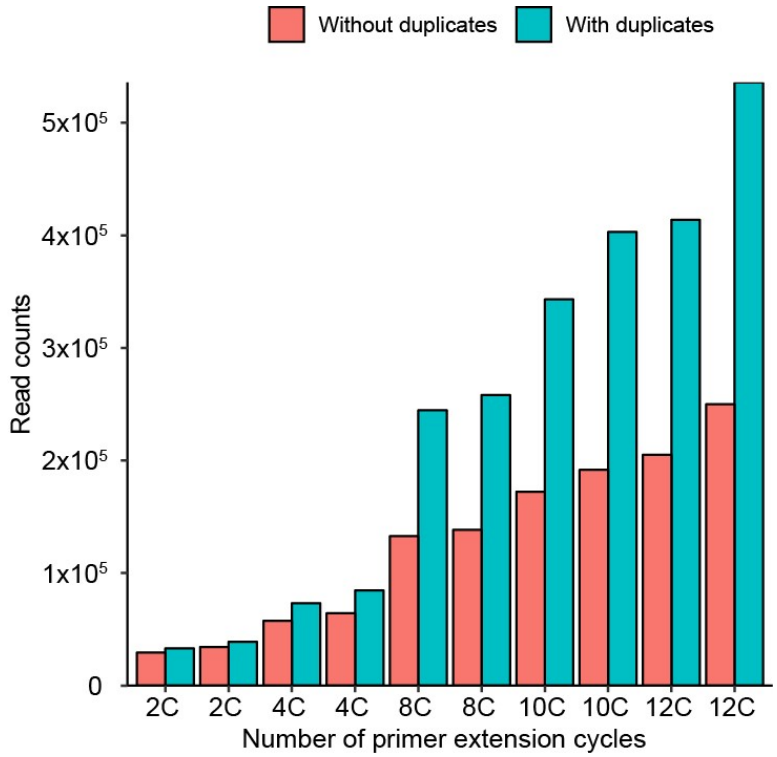


Figure 4-18: Increasing read output with increasing number of primer extension cycles.
Each primer extension cycles contains two technical replicates. Libraries were prepared from the spleen of a 24-month-old C57BL/6 mouse by Dan Bolland. The output reads from the deduplication pipeline were merged with low UMI count reads and additionally deduplicated based on the entire read sequence (without anchor-UMI).

J1 primer Identical sequence Anchor UMI

```
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGGTCCTGTTGCTA
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGGCAGGCATTCCC
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGTCTGGTCATCAG
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGGCGTTGGCCAGC
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGTCCCCTACCCGCAGATCGAGGAGCAGCCACGCAAGGCC
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGGATCGGGCTCCC
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGGCAAGCGGGCAC
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGTAAGGATTCTAC
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGCCTGGTCTTCAA
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGTCAGGTCGCTCA
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGTGTCAACCCTTC
CCCTGTGCCCCAGACATCGAAGTACCAGTGAGGG ... TCCGCATAATGTAGGAGCAGCTCCACGCGTCC
```

Figure 4-19: Presence of two anchor sequences within a single read.
The design of the anchor-UMI sequence does not allow a second anchor-UMI sequence to be ligated on. A template switching event could explain the appearance of these chimeric reads. All the read sequences shown here have identical sequences apart from the UMI. The J primer is highlighted in red with the 5 bp sequence used for mispriming correction shown in grey. The anchor sequences are highlighted in green.

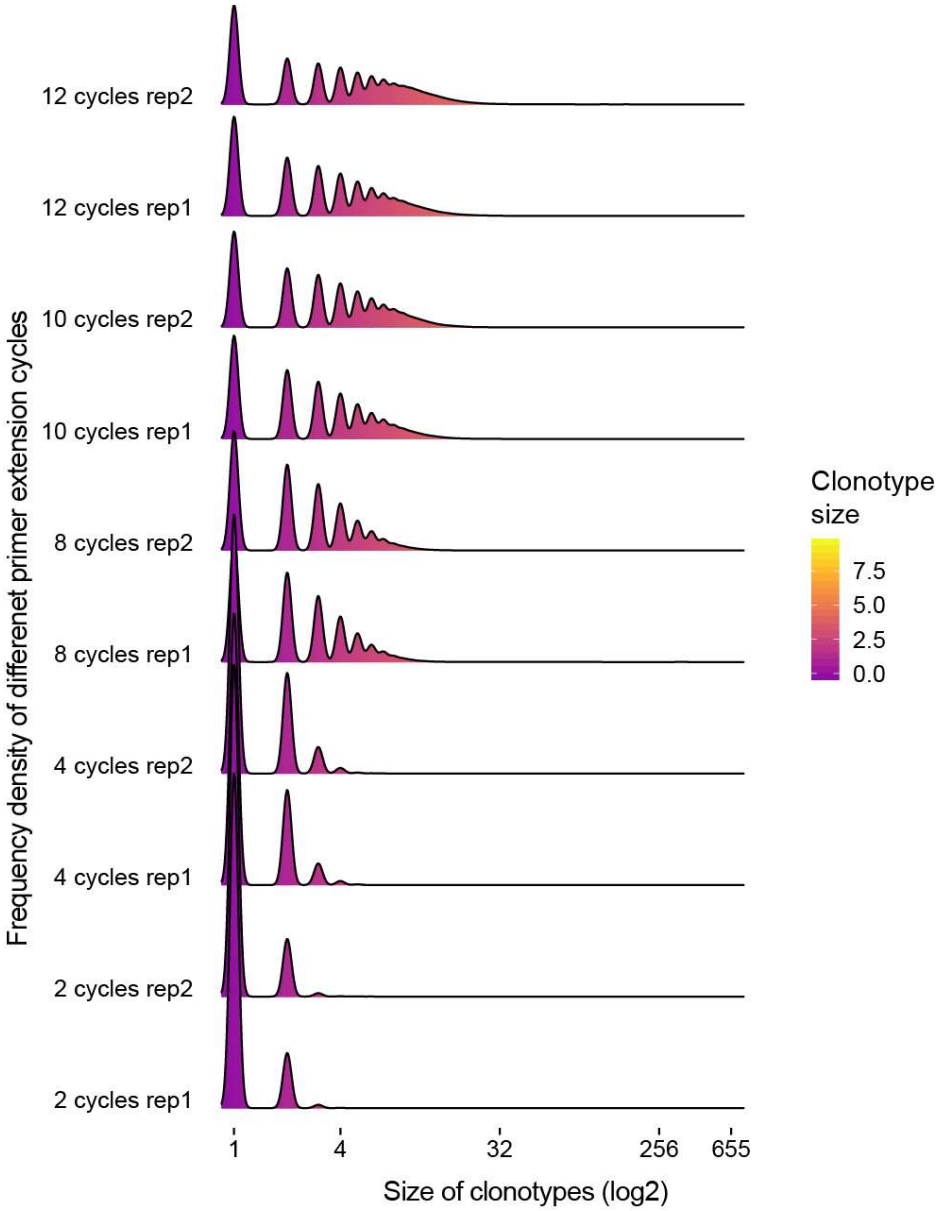


Figure 4-20: The size of clonotype groups increases with increasing primer extension cycle numbers.
Because the samples came from an unimmunised mouse, the number of clones is expected to be low. In addition, all samples represent technical replicates, which are expected to yield the same clonotype sizes. Plotted with the R package ggridges v0.5.0.

4.4.4 Deduplication and annotation of short reads with Partis

IgBlast annotation of recombined genes works by first searching for the V gene and subsequently masking the sequence that matched the top germline V gene (Ye et al., 2013). The masking is performed to avoid incorrect hits when searching for the D and J genes. After a V gene has been successfully found, the search for the J gene follows. The D gene is the last to be searched for and is assumed to be positioned between the V and J genes. The short reads inherently produce a low quality V gene match, which normally results in the removal of these reads. However, higher confidence calls can be independently obtained from the V end reads and because the aim is to only obtain the V gene

identity, a single V end sequence from a UMI group is sufficient. This meant that the V end reads never went through deduplication and consensus sequence construction. The current annotation of short reads is not ideal as IgBlast initially searches for the short V sequence within the J end reads, which could produce inaccurate junction boundaries and perhaps even impact the D and J genes calls. Previous iterations of the pipeline used a separate script to fill in the missing sequence between the J end and V end reads to allow the use of annotation tools such as IMGT/HighV-Quest. However, by joining the V and J reads together and reverse complementing th sequence so they are in the VDJ orientation I was able to use the annotation tool partis (Ralph and Matsen, 2016), which is capable of filling in potential deletion within sequences (Figure 4-21). As a result, the newest version of the pipeline deduplicated the V end reads in the same manner as the J end reads and creates high quality consensus sequences around which gaps can be filled.

Figure 4-21: Partis asnnotation example output.
Partis marks the missing sequence between the V and J reads as a deletion and fills it with germline sequence. The sequence between the red bases marking cytosine (TGT) and tryptophan (TGG) is the CDR3.

4.5 Commands and additional output of BabrahamLinkON

The following section contains documentation of the BabrahamLinkON commands and provides examples of additional plots optionally produce during processing, which can be used to verify everything has run as expected or can be helpful for troubleshooting.

4.5.1 Preclean commands and output documentation

All input fastq files are assumed to have been adaptor and quality trimmed. Adaptor and low-quality end sequence trimming can be performed using Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which is wrapper script for the adapter-trimming tool Cutadapt (Martin, 2011). For Igκ, the low diversity of the first four nucleotides of the J end R2 (read 2) results in low quality bases that should be additionally trimmed.

BabrahamLinkON preclean - choosing pipeline	
`-h, --help`	show the help message and exits
`--version`	show the program's version number and exit
Choose pipeline:	
`umi`, `short`, `short_anchor`,	Four options are available depending on the reads. Reads that can be assembled and have an anchor-UMI are used with the `umi` option. Reads that are too short to assemble and do not have an anchor-UMI

mispriming_error	are used with the short option. Reads that are too short to assemble but contain an anchor-UMI are used with the short_anchor option. Finally, a method to estimate the inaccuracy of mispriming correction based on germline reads is provided with the mispriming_error option.

Arguments available in precleaning	
Arguments shared by all:	
-h, --help	show the help message and exits
--species SPECIES	Which species or locus: mmu – *Mus musculus*; hsa - *Homo sapiens*; mmuk – *Mus musculus* kappa). So far, only these three options are available. [mmu]
-t NTHREADS, --threads NTHREADS	Number of threads to use. [1]
-v INPUT_V,--V_r1 INPUT_V	Input fastq file with V end sequences
-j INPUT_J, --J_r2 INPUT_J	Input fastq file with J end sequences
--fast	Perform fast inaccurate J identification (additionally need to use --no_mispriming). Only recommended if mispriming correction is not desired.
--no_mispriming	Don't perform mispriming correction. (Not recommended)
--prefix PREFIX	Prefix of the output file. By default, the prefix is the basename of the V end fastq file.
--out OUT_DIR	Output directory. By default a new folder is created with the V end fastq file basename in the same directory as the fastq files. In cases were users do not have write permission in the current directory, a custom output directory can be specified.
--verbose	Print detailed progress
--plot	Plot alignments of reads to the germline J genes. Useful as a quick check for the consistency of germline J gene levels observed between samples.
-q Q_SCORE, --q_score Q_SCORE	Minimum Phred quality score allowed for bases in UMI. This is to guarantee that only high quality UMIs are used. [30]
-ul UMI_LEN, --umi_len UMI_LEN	Length of the UMI [12]
--j_len J_LEN	Length of J end sequence, [50] bp into read (--in_len), to add to the UMI to further diversify the UMI. [0]
--in_len IN_LEN	How many bps to go into the J end sequence before taking --j_len to increase diversity of the UMI. [50]
--an1 AN1	Adaptor 1 anchor sequence. This is only used if custom anchor sequences were designed. [GACTCGT]
--an2 AN2	Adaptor 2 anchor sequence. This is only used if custom anchor sequences were designed. [CTGCTCCT]
--keep_germline	Skip germline removal step. Useful if the deduplication and analysis of germline sequences captures is desired.
--keep_pear	Do not delete output files from pear. Useful for troubleshooting assembly of paired-end reads.

Arguments unique to the `UMI` and `short_anchor` pipeline:	
`-ba BEYOND_ANCHOR, --beyond_anchor BEYOND_ANCHOR`	Length of V end to take beyond the anchor sequence. Useful if the UMI is short (e.g. 6 bp) and part of the V region that contains sonication variability is required to increase UMI diversity.
Arguments unique to the `short` and `short_anchor` pipeline:	
`--ref REF_PATH`	Igh reference files path

Available command options for each pipeline

BabrahamLinkON preclean `UMI` pipeline

```
preclean.py umi     [-h][--species SPECIES][-t NTHREADS]
                    [-v INPUT_V][-j INPUT_J][--fast]
                    [--no_mispriming][--prefix PREFIX]
                    [--out OUT_DIR][--verbose][--plot]
                    [-q Q_SCORE][-ul UMI_LEN][--j_len J_LEN]
                    [--in_len IN_LEN][--an1 AN1][--an2 AN2]
                    [--keep_germline][--keep_pear]
                    [-ba BEYOND_ANCHOR]
```

BabrahamLinkON preclean `short` pipeline

```
preclean.py short   [-h][--species SPECIES][-t NTHREADS]
                    [-v INPUT_V][-j INPUT_J][--fast]
                    [--no_mispriming][--prefix PREFIX]
                    [--out OUT_DIR][--verbose][--plot]
                    [-q Q_SCORE][-ul UMI_LEN][--j_len J_LEN]
                    [--in_len IN_LEN][--keep_germline]
                    [--keep_pear][--ref REF_PATH]
```

BabrahamLinkON preclean `short_anchor` pipeline

```
preclean.py short_anchor
                    [-h][--species SPECIES][-t NTHREADS]
                    [-v INPUT_V][-j INPUT_J][--fast]
                    [--no_mispriming][--prefix PREFIX]
                    [--out OUT_DIR][--verbose][--plot]
                    [-q Q_SCORE][-ul UMI_LEN][--j_len J_LEN]
                    [--in_len IN_LEN][--an1 AN1][--an2 AN2]
                    [--keep_germline][--keep_pear][--ref REF_PATH]
                    [-ba BEYOND_ANCHOR]
```

Precleaning of long reads with UMIs can be run with the following example command:

```
preclean.py umi -v <R1.fastq.gz> -j <R2.fastq.gz> --species mmu
--threads 8 --q_score 30 --umi_len 12
```

where `umi` selects the long read with UMI option of the pipeline. The <`R1.fastq.gz`> is the location of the V end fastq file, while the <`R2.fastq.gz`> is the location of the J end fastq file.

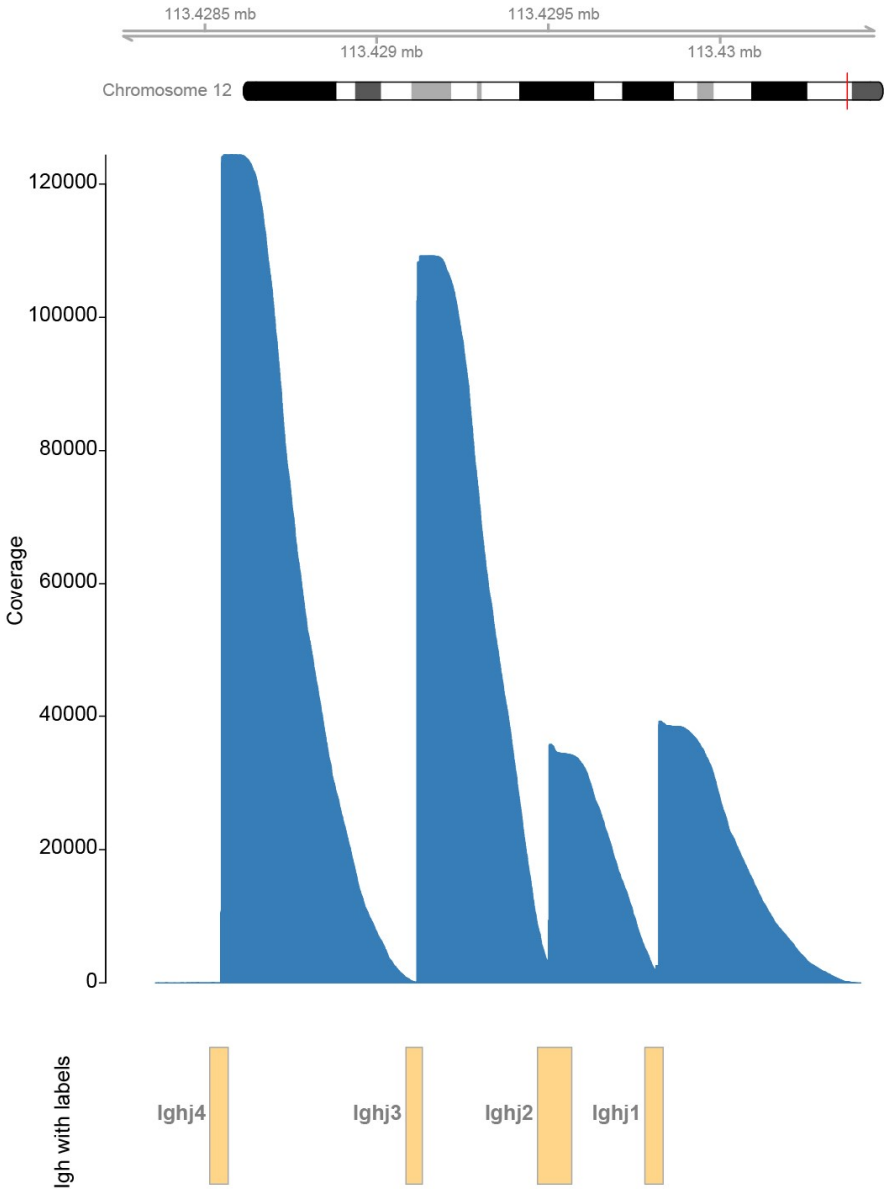


Figure 4-22: Coverage plot of germline VDJ-seq reads over the mouse J genes.
The plot is optionally produced from the germline reads that are removed during precleaning. The sigmoid curve on the right-hand side of each peak is characteristic of the variable sonication.

The germline read coverage plots (Figure 4-22) are produced using the Gviz package v2.32.1+ (Hahne and Ivanek, 2016). Genomic intervals are extracted from a Samtools v1+ (Li et al., 2009) sorted and indexed BAM file using the GenomicsRanges package v1.28.6+ (Lawrence et al., 2013) and the track annotation is extracted from Ensembl using the biomaRt package v2.32.1+ (Durinck et al., 2009). All of the R packages are part of Bioconductor (Huber et al., 2015). The germline VDJ-seq read coverage plot can be used to qualitatively compare the germline reads captured between samples. The visible alignment of the reads beyond the J genes confirms their germline identity. The plot enables a quick validation that the germline filter is correctly removing germline reads.

BabrahamLinkON preclean `mispriming_error` tool	
`preclean.py mispriming_error [-h] [--species SPECIES] [-t NTHREADS] --input_dir IN_DIR`	
Arguments:	
`-h, --help`	show the help message and exits
`--species SPECIES`	Which species or locus: `mmu` – *Mus musculus*; `hsa` - *Homo sapiens*; `mmuk` – *Mus musculus* kappa). So far, only these three options are available. [`mmu`]
`-t NTHREADS, --threads NTHREADS`	Number of threads to use. [`1`]
`--input_dir IN_DIR`	Input directory that was created by the preclean.py pipelines. This folder needs to contain the germline fastq files (not compatible with `--keep_germline`).

4.5.2 Deduplication commands and output documentation

BabrahamLinkON Deduplicate choosing pipeline	
`-h, --help`	show the help message and exits
`--version`	show the program's version number and exit
Choose pipeline:	
`umi`, `short`, `short_anchor`, `no_anchor`, `umi_seq_logo`, `reverse_complement`	There are six available options depending on the reads. Reads that have been assembled and have been split into two anchor fastq files with the UMI withing the read name are used with the `umi` option. Reads that were too short to assemble and do not have an anchor-UMI are used with the `short` option. Reads that were too short to assemble but contain an anchor-UMI are used with the `short_anchor` option. For long reads that have been assembled but do not contain an anchor-UMI are used with the `no_anchor` option. For creating sequence logos of the UMI the `umi_seq_logo` option can be used. Finally, for reverse complementing the output sequences from the JDV to the VDJ orientation that is required for some annotation tools, the `reverse_complement` option can be used.

The `no_anchor` option is a special use case as there are no explicit commands in the precleaning stage of the pipeline to prepare these sequences for deduplication. The `umi` precleaning pipeline can be used on these sequences as it is designed for long reads. However, because the sequences do not contain an anchor sequence it will fail to identify it, but it will still extract x amount of nucleotides from the V end that can be used as a UMI for deduplication. Unfortunately, in order to pass the sequences to the deduplication pipeline the file will need to be manually renamed from `` `_other_J` `` suffix to the `` `_all_jv` `` suffix. The deduplication pipeline looks for files with unique suffixes for each pipeline. This was implemented as a precaution to avoid using the wrong input files in downstream steps of the pipeline.

Arguments available in deduplicate	
Arguments shared by all:	
`-h, --help`	Show the help message and exits
`--input_dir IN_DIR`	Input directory that was created from precleaning.
`--out OUT_DIR`	Output directory. By default, a folder is created in main directory.
`-t NTHREADS, --threads NTHREADS`	Number of threads to use for processes that can be performed in parallel. [`1`]
`--mismatch MISMATCH`	Number of mismatches allowed between the consensus sequence and the sequences that were used to create the consensus. [5]
`--min_reads MINREADS`	Minimum number of reads in UMI group, if less than or equal to [2] then these reads will be written into a separate file.
`--gt_ratio GTRATIO`	Ratio of good to total reads within a UMI group/stack. A read is classified as good if it has less than the allowed number of mismatches (see `--mismatch`). This allows the identification of UMI group with too many sequence error or with incorrectly grouped reads. The accepted values range from 0 to 1. [1]
`--stats`	Output stats from UMI deduplication [False]
`--umi_correction`	Perform correction of errors that might be present within the UMI. Very slow for large samples, but can be used with downsampled files to identify optimal length of V end for UMI.
`--threshold THRESHOLD`	Number of mismatches allowed in UMI when doing UMI correction. [1]
`--skip_unclear`	Do not process unclear J reads (see mispriming correction). [False]
`--keep_mh`	Keep multiple hit J reads (see mispriming correction). [False]
`--use_j`	Deduplicate using J identity from mispriming correction.
`--ignore_umi`	Deduplicate without using the UMI (for troubleshooting or QC).
`--no_consensus`	Do not output consensus sequence, but instead only return the first sequence in the UMI stack (for troubleshooting or QC).
`--j_trim`	Trim J primer when comparing sequences to the consensus sequence. This allows misprimed J sequences to be corrected by the more frequent true J sequence. The length of the sequence to be trimmed should be set to the longest J gene. [25]
`--no_msa`	Do not use multiple sequence alignment (KAlign2) to derive consensus sequence. Recommended only for troubleshooting and QC purposes. [False]
`--fq`	Output fastq instead of fasta. Some annotation tools accept fastq files instead of fasta.
`--cons_no_qual`	Make consensus sequence without using fastq quality scores.
`--with_N`	Output consensus sequences with ambigious N bases. If the consensus sequence contains N bases (ambiguous with tied quality scores), normally the sequence that best matches the consensus is used to fill in the N bases. This is done to avoid problems with interpretation of N by annotation tools.
Arguments unique to `umi` and `short_anchor` pipeline	
`--an1 AN1`	Default: GACTCGT
`--an2 AN2`	Default: CTGCTCCT

Available command options for each pipeline

BabrahamLinkON deduplicate `UMI` pipeline

```
deduplicate.py umi
                    [-h] --input_dir IN_DIR [--out OUT_DIR]
                    [--threads NTHREADS][--mismatch MISMATCH]
                    [--min_reads MINREADS][--gt_ratio GTRATIO]
                    [--stats][--umi_correction]
                    [--threshold THRESHOLD][--skip_unclear]
                    [--keep_mh][--use_j][--ignore_umi]
                    [--no_consensus][--j_trim][--no_msa][--fq]
                    [--cons_no_qual][--with_N][--an1 AN1]
                    [--an2 AN2]
```

BabrahamLinkON deduplicate `short` pipeline

```
deduplicate.py short
                    [-h] --input_dir IN_DIR [--out OUT_DIR]
                    [--threads NTHREADS][--mismatch MISMATCH]
                    [--min_reads MINREADS][--gt_ratio GTRATIO]
                    [--stats][--umi_correction]
                    [--threshold THRESHOLD][--skip_unclear]
                    [--keep_mh][--use_j][--ignore_umi]
                    [--no_consensus][--j_trim][--no_msa][--fq]
                    [--cons_no_qual][--with_N]
```

BabrahamLinkON deduplicate `short_anchor` pipeline

```
deduplicate.py short_anchor
                    [-h] --input_dir IN_DIR [--out OUT_DIR]
                    [--threads NTHREADS][--mismatch MISMATCH]
                    [--min_reads MINREADS][--gt_ratio GTRATIO]
                    [--stats][--umi_correction]
                    [--threshold THRESHOLD][--skip_unclear]
                    [--keep_mh][--use_j][--ignore_umi]
                    [--no_consensus][--j_trim][--no_msa][--fq]
                    [--cons_no_qual][--with_N][--an1 AN1]
                    [--an2 AN2]
```

BabrahamLinkON deduplicate `no_anchor` pipeline

```
deduplicate.py no_anchor
                    [-h] --input_dir IN_DIR [--out OUT_DIR]
                    [--threads NTHREADS][--mismatch MISMATCH]
                    [--min_reads MINREADS][--gt_ratio GTRATIO]
                    [--stats][--umi_correction]
                    [--threshold THRESHOLD][--skip_unclear]
                    [--keep_mh][--use_j][--ignore_umi]
                    [--no_consensus][--j_trim][--no_msa][--fq]
                    [--cons_no_qual][--with_N]
```

Deduplication of long reads with UMIs can be run with the following example command:

```
deduplicate.py umi --input_dir <directory_path> --threads 8 --stats
--min_reads 2 --gt_ratio 1 --mismatch 5
```

where `umi` selects the long read with UMI option of the pipeline. The `<directory_path>` specifies the path of the directory created during precleaning that contains a fastq file for each group of anchor sequences.

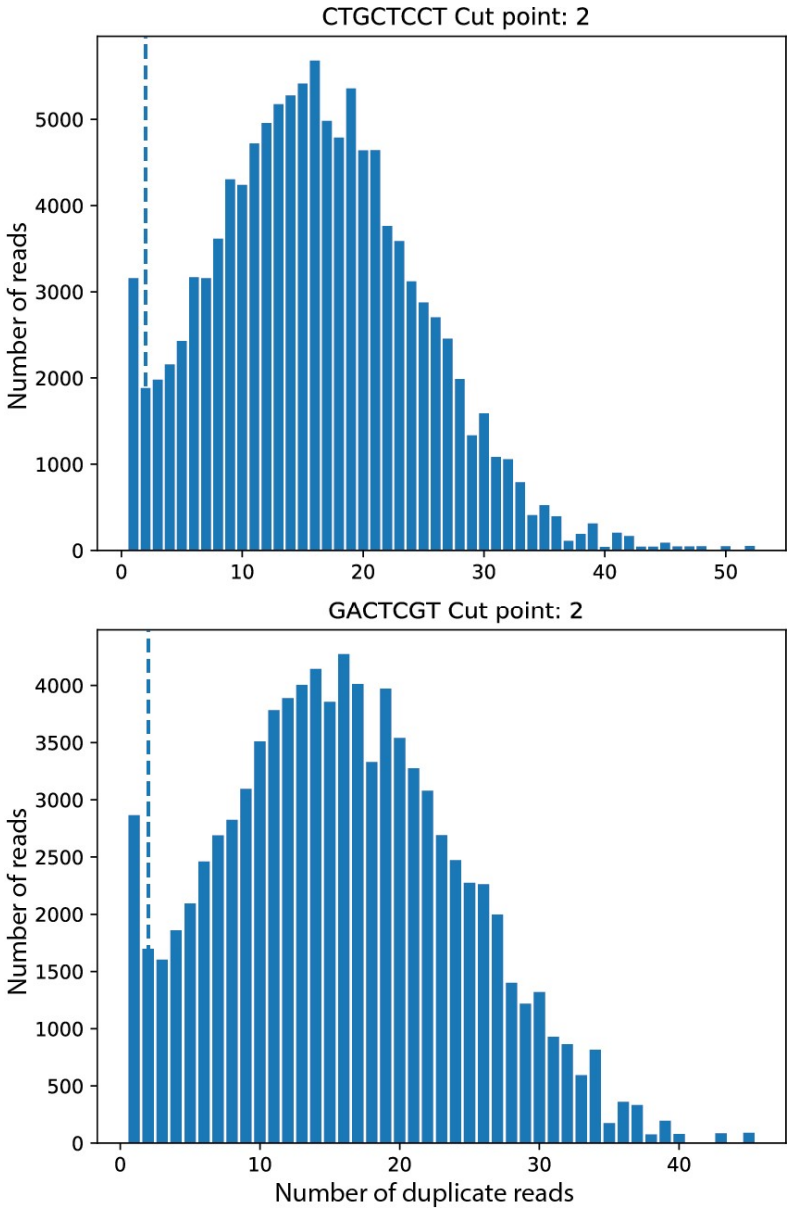


Figure 4-23: Duplicate histograms are output as part of the deduplication pipeline.
The dashed line corresponds to the low count UMI groups cut off point. Anything equal or less than the cut off will be placed into a separate file. This plot is useful to determine if the sample has sufficient over sequencing and if the cut-off point is reasonable. Histograms with only the left tail cut-off are indicative of sufficient over sequencing. In cases where only the right-hand tail is visible, indicative of under sequencing, the choice of including the low count UMI groups in downstream analysis can be made.

Table 4-10: Log of duplication level/count of each UMI pre and post deduplication.
The duplication values can be found in the <sample_name>_umi_per_position.tsv file created by the deduplication pipeline. The lower counts post deduplication correspond to the reads not passing the good-to-total ratio filter. The duplication counts can be used to create an equivalent histogram to the one shown in Figure 4-23. Counts between 6 and 21 have been omitted here.

Counts	Instances_pre	Instances_post
1	5062	5062
2	2332	2253
3	1864	1798
4	1747	1679
5	1650	1584
6	1487	1433
…	…	…
21	31	30
22	12	11
23	9	8
24	6	6
25	4	4
26	1	1
27	1	1
28	2	2
30	1	1

Table 4-11: Log of each UMI group/stack pre and post deduplication.
The information of each UMI stack can be found in the <sample_name>_per_umi_per_position.tsv file created by the deduplication pipeline. The log contains information on the total number of reads observed for each UMI (total counts). It also shows the number of mismatches/differences each sequence contains when compared to the consensus sequence, and provide the MSA output for unique sequences in a comma delimited fasta format. The alignment output is very useful in troubleshooting cases were high number of sequences are removed with the good-to-total ratio filter. The number of times each UMI was observed is also shown, but this metric is only useful when performing UMI correction where observations higher than one indicate merging of two UMI groups.

UMI	Total counts pre	Times observed pre	Total counts post	Times observed post	Alignments	Consensus differences
AAAAAATAGCAC	4	1	4	1	>0_AAAAAATAGCAC_1, AGTG…ACCA, >1_AAAAAATAGCAC_3, AGTG…ACCA	1,0,0,0
AAAAACCATAGA	1	1	1	1	>0_AAAAACCATAGA_1, AGTG…GTTG	0
AAAAACCCATCT	1	1	1	1	>0_AAAAACCCATCT_1, CCCT…CCCA	0
AAAAAGCGTGCT	5	1	5	1	>0_AAAAAGCGTGCT_5, AGTG…TGAG	0,0,0,0,0
AAAAAGGGACTT	2	1	2	1	>0_AAAAAGGGACTT_1, CCCT…TGTT, >1_AAAAAGGGACTT_1, CCCT…TGTT	0,2
AAAAAGTTAGAG	5	1	5	1	>0_AAAAAGTTAGAG_5, CCCT…CATT	0,0,0,0,0
…	…	…	…	…	…	…

High number of sequences being filtered out by the good-to-total ratio filter is usually a sign that either something has gone wrong with the library preparation or the sequences are not being correctly grouped. A very useful output of the deduplication pipeline for the troubleshooting of such cases is the per UMI log. A portion of the log is shown in Table 4-11, which shows the number of differences/mismatches of each sequence from the consensus and provides the MSA output for each unique sequence. The consensus difference highlights outlier sequences that may contain PCR and sequencing errors or have been incorrectly grouped. Incorrect grouping can arise from low diversity UMIs, template switching during primer extension or during UMI error correction where lower count UMI groups are merged with higher count UMI groups that differ in their UMI sequence by only 1 hamming distance. If higher numbers of mismatches are present within the UMI group sequences, the number of allowed differences between a UMI stack sequence and the UMI stack consensus sequence can be changed from the default 5.

BabrahamLinkON deduplicate `umi_seq_logo` **tool**	
`deduplicate.py umi_seq_logo [-h] --input_dir IN_DIR [--out OUT_DIR] [--an1 AN1][--an2 AN2]`	
Arguments:	
`-h, --help`	show the help message and exits
`--input_dir IN_DIR`	Input directory that was created from precleaning.
`--out OUT_DIR`	Output directory. By default, a deduplicated folder is created in main directory, or if the folder already exist the output is written into the folder.
`--an1 AN1`	Anchor 1 sequence [GACTCGT]
`--an2 AN2`	Anchor 2 sequence [CTGCTCCT]

The weblogo package v3.5.0 was used as the backend for the generation of sequence logo plots of the UMI (Crooks et al., 2004). An example sequence plot is shown in Figure 4-24. Such plots can be used as a troubleshooting tool to ensure an equal representation of all the bases within the UMI. This can be especially useful is cases were a proxy UMI, such as the V end sonication variability or a portion of the junction sequence, is being used.

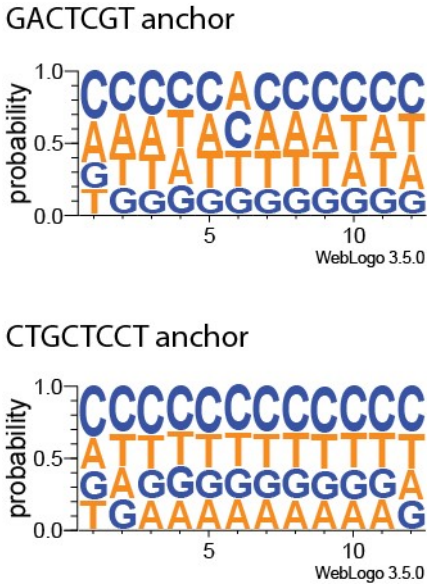


Figure 4-24: An example sequence logo plot of the UMI sequence.
A separate plot is generated for each of the anchor sequences. These plots are useful for troubleshooting overrepresentation of bases at certain positions of the UMI. This is especially useful when using V end or J end sequences as a proxy for a true UMI, where an equal distribution may not be present.

BabrahamLinkON deduplicate `reverse_complement` **tool**	
`deduplicate.py reverse_complement [-h] [--an1 AN1] [--an2 AN2] --input INPUT [--fq]`	
Arguments:	
`-h, --help`	show the help message and exits
`--an1 AN1`	Anchor 1 sequence [GACTCGT]
`--an2 AN2`	Anchor 2 sequence [CTGCTCCT]
`--input INPUT`	Input file or directory with files. By default, a fasta file is expected.
`--fq`	Instead of a fasta, the input and the output files are fastq.

4.5.3 Annotation and assembly command documentation

BabrahamLinkON annotation and assembly - choosing pipeline	
`-h, --help`	show the help message and exits
`--version`	Show the program's version number and exit
Choose pipeline:	
`umi, short, assemble_only`	Two options are available depending on the reads. At this stage only the state of the reads, whether they are assembled into a single read or still split into J end and V end reads matters. Assembled reads are used with the `umi` option. For short reads, the `short` option needs to be used. Short read annotation is performed separately for each read end, as the J end may not contain enough V sequence to obtain a high confidence call. The end with the highest score is used for the final gene call. As an additional control, the V gene identified in the

	precleaning stage with Bowtie2 alignment can be manually compared with calls obtained from IgBlast. The third `assemble_only` option is used if annotation has already been performed and only the clone assembly is required. Clone assembly can take a long time on very large sample, and therefore in some cases were only the annotation of the V(D)J genes is desired then assembly is often skipped. In addition, if the clonal overlap between multiple samples is desired, the tsv files of all the samples can be supplied and a single merged output tsv will be produced.

Arguments available in annotation and assembly	
Arguments shared by all:	
`-h, --help`	Show the help message and exits
`--threshold THRES`	Number of differences allowed between CDR3 sequences [1]
`--only_v`	Use only V identity and CDR3 for clone assembly
`--out OUT_DIR`	Output directory. By default, the files are written into the directory created during deduplication.
Arguments unique to the `umi` and `short` pipeline	
`-fa FASTA, --fasta FASTA`	Input fasta file from deduplication.
`--plot`	Plot V and J scores with cutoff.
`--full_name`	Retain full name of first V and J genes, otherwise the allele information is removed.
`--minimal`	Work with and output only a minimal table
`--threads NTHREADS`	Number of threads to use [1]
`--species SPECIES`	Which species (mmu hsa mmuk) [mmu]
`--aux AUX`	Custom aux file required by Igblast. The default aux files have been included with the pipeline installation.
`--custom_ref`	Use AEC custom reference for Igblast.
`--v_cutoff V_CUTOFF`	IgBlast V_SCORE cutoff [>50]
`--j_cutoff J_CUTOFF`	IgBlast J_SCORE cutoff [>35]
`--skip_assembly`	Do not assemble clones into clonotypes
`--call_dj`	Call DJ recombination alongside VDJ.
Arguments unique to the `short` pipeline:	
`-fq V_FASTQ, --v_fastq V_FASTQ`	V end R1 fastq file
Arguments unique to the `assemble_only` pipeline:	
`-tsv TSV_FILES, --tsv_files TSV_FILES`	Input tsv files from previous runs of `assemble_clones`

Annotation and clone assemble of long reads with UMIs can be run with the following example command:

```
assemble_clones.py umi –fa <input_fasta> --threads 8 species
mmu --full_name –-plot
```

where `umi` selects the long read with UMI option of the pipeline. The `<input_fasta>` specifies the path of the fasta file produced by the deduplication.

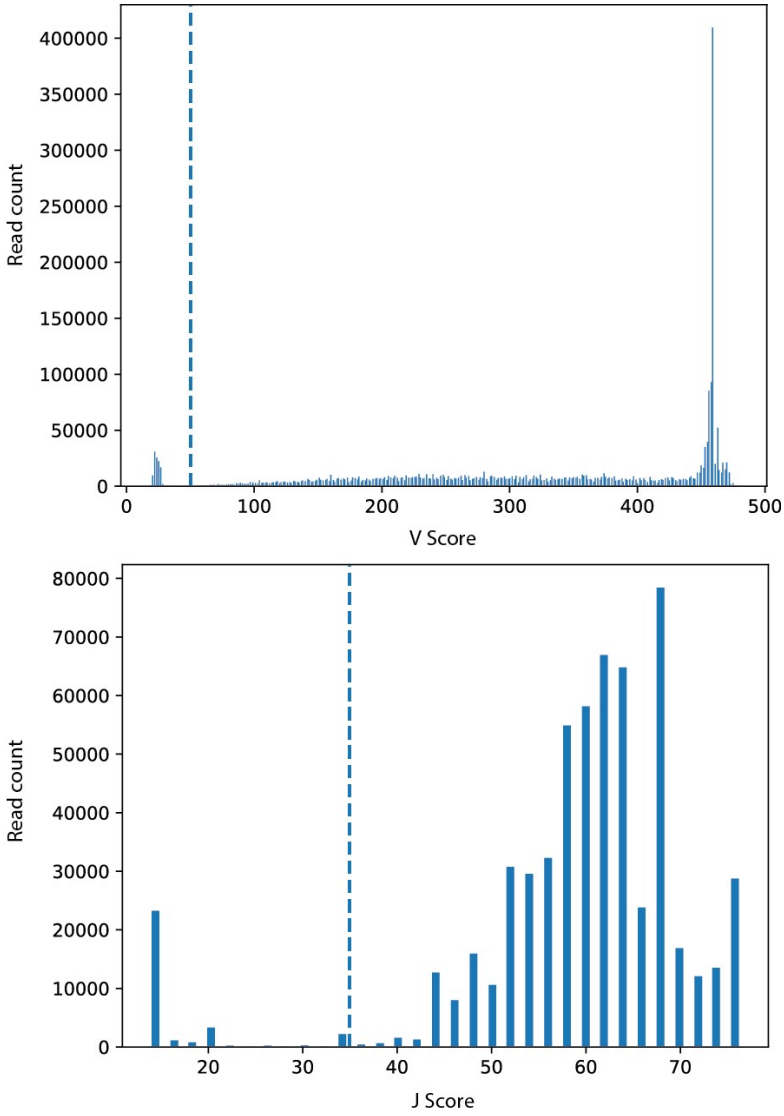


Figure 4-25: Histograms of V and J scores from IgBlast.
The vertical dashed line marks the score filter cut-off. The majority of reads have a high score.

The histogram of V and J scores can be optionally produced during the annotation stage of the pipeline (Figure 4-25). It can be used to verify that most sequences contain high quality scores and that the default threshold is a reasonable cut-off. The low scores tend to be DJ and germline reads.

4.6 Discussion

The development of a comprehensive analysis pipeline for VDJ-seq, along with a desire to profile low cell number samples, led to the optimisations and improvement of the VDJ-seq protocol. The systematic correction of mispriming events along with the incorporation of a high diversity UMI allowed the profiling of antigen receptor repertoires at the level of individual clones. Using multiple reads harbouring identical UMIs enabled the correction of PCR and sequencing errors and the filtering of aberrant reads. This opens up VDJ-seq for potential clinical applications such as studies of clonal expansion in leukaemia, along with basic research of repertoires in mouse models and in human health and disease (Galson et al., 2014; Hou et al., 2016b; Robinson, 2015). VDJ-seq is also able to capture substantially higher numbers of clonotypes compared to other gDNA-based methods and produce an unbiased quantification of all V genes (Chovanec et al., 2018). By additionally creating an annotation reference for recombined D genes, it is now possible to quantify the full range of VDJ, DJ and unrecombined sequences captured with VDJ-seq, which will enable more quantitative studies of the earliest recombination events in cells such as CLPs.

The field of antigen receptor (AgR) repertoire studies has seen immense progress since the first adaptations of techniques for next generation sequencing in 2009 (Weinstein et al., 2009). The challenges of developing high-throughput wet lab methodologies that can capture the AgR repertoire without bias face an equally great challenge in the analysis of the data (Chaudhary and Wesemann, 2018; Greiff et al., 2015; Yaari and Kleinstein, 2015). Some of the future developments to improve VDJ-seq should address the large proportion of germline fragments that soak up the sequencing depth. The depletion of abundant sequences by hybridisation (DASH) method utilises CRISPR (clustered regularly interspaced short palindromic repeats) with the associated Cas9 nuclease and guide RNAs to specifically cleave undesired sequences within an Illumina library just before PCR amplification (Gu et al., 2016). The cleavage prevents the amplification of these sequences, effectively depleting them from the final library. Using guide RNAs for the germline reads within a VDJ-seq library should make it possible to efficiently eliminate them. The advantage of DASH is that it can be used with small amounts of starting material and without losing substantial amounts of non-targeted sequences. The elimination of germline fragments would substantially reduce sequencing costs and perhaps permit the sequencing of multiple technical replicates that would lead to the capture of a greater portion of the repertoire.

The theoretical diversity of AgR repertoires makes the probability of individuals sharing clones (public clones) unlikely. However, several studies have observed the presence of public clones and pointed to genetic background and antigen driven convergence of BCRs as the mechanisms of public clone emergence (DeKosky et al., 2016; Greiff et al., 2017a; Jackson et al., 2013). Interestingly, public clones

can be predicted with high accuracy based on features of the CDR3 region (Greiff et al., 2017b). The ability to capture both the heavy and light chains from the same cell has additionally revealed that Igl public sequences are more common than Igh public sequences, while the knowledge of the recombination status of both loci can additionally provide invaluable insight into basic mechanisms such as allelic inclusion and to a more broader understanding of diseases (DeKosky et al., 2013, 2016). In addition, the ability to produce and screen mAbs with native pairing, as opposed to random pairing of current antibody screening libraries, provides a massive resource for future therapeutics development (Jayaram et al., 2012; Wang et al., 2018). As a future direction, merging the low cost of VDJ-seq with the ability to examine Igh:Igl pairing is foreseeable with the recent advance in split-pool barcoding techniques (Cao et al., 2017; Rosenberg et al., 2018; Svensson et al., 2018). A split-pool barcoding adaptation for AgR-seq could soon enable the analysis of millions of paired recombination events at low cost and without the need for specialised equipment.

5 Promoter interaction dynamics in a model system of human embryonic development

My interests lie in understanding how genome conformation regulates B cell development. Our group has recently published a highly collaborative study in which we have performed promoter capture Hi-C (PCHiC) on mouse pre-B cells (Koohy et al., 2018). The ultimate aim of developing techniques for the visualisation of entire PCHiC datasets is to be able to investigate genome organisation changes taking place during the transition from pro-B to pre-B cells. Unfortunately, current requirements for large quantities of starting cells prohibit the generation of datasets from pro-B cells. A complementary project in progress in the lab aims to overcome this limitation by refining PCHiC for use with substantially smaller numbers of cells. My contribution to this project has been to develop more informative ways of analysing these complex datasets. For this study, we availed of a collaborative opportunity to develop this platform in collaboration with Peter Rugg-Gunn's group, who have developed the important and tractable human naïve and primed hPSCs system from which we were able to study high quality PCHiC and related datasets.

5.1 Background

Pluripotent stem cells have the capacity to from all cell lineages of the adult body. The derivation of mouse embryonic stem cells (mESCs) from the entire mouse blastocyst, or inner cell mass (ICM), led to the stabilisation of their *in vivo* transient state of pluripotency and was instrumental in expanding our understanding of cell fate decisions and the ability of these cells to differentiate into numerous cell types under specific growth conditions. These studies were later followed by the derivation of human pluripotent stem cells (hPSCs) from blastocyst stage embryos acquired by *in vitro* fertilisation (IVF) techniques (Thomson et al., 1998), many of which are still commonly used today (Guhr et al., 2018). Being able to maintain hPSCs in culture indefinitely, while retaining their capacity to differentiate, provides an indispensable tool for expanding our knowledge of human embryonic development and disease modelling with the ultimate goal of harnessing the plasticity of these cells for regenerative medicine. However, *in vitro* hPSCs of early studies displayed molecular and morphological differences to their mouse counterparts, despite the derivation being performed from an equivalent preimplantation embryonic developmental stage (Collier and Rugg-Gunn, 2018). It was later discovered that the culture conditions, used for deriving and maintaining human preimplantation epiblast cells, were incapable of retaining them in a preimplantation-like state, and instead stabilised them in a later developmental stage characteristic of postimplantation epiblast cells (Nakamura et al., 2016).

The maintenance of mESC in different culture media formulations leads to morphological, gene expression and epigenetic differences between the cells, while still allowing them to retain their

pluripotent characteristics. mESCs grown in the presence of leukaemia inhibitory factor (LIF) and 2i (two inhibitors of MEK and GSK3) gave rise to a homogeneous population of cells referred to being in a 'naïve' state of pluripotency, representative of preimplantation epiblast cells (Ying et al., 2008). Alternatively, culturing mouse ESC in the presence of FGF and Activin A generates a cell type that is representative of postimplantation epiblast cells (mEpiSCs), which are classified as being in a 'primed' state of pluripotency (Brons et al., 2007; Tesar et al., 2007). Molecular and phenotypic analysis of naïve mESC and primed mEpiSCs reveals that hPSCs more closely resemble primed mEpiSCs. In the past few years, considerable progress has been made towards formulating culture conditions and reprogramming methods for deriving naive hPSCs from primed hPSCs and maintaining them in culture. One of two most prominent methods for reprogramming and maintaining primed hPSCs in a naïve state is using the 5i/L/A media formulation, which utilises five inhibitors along with LIF and Activin A (Theunissen et al., 2014). The second method employs the transient induction of pluripotency transcription factors to reprogram primed cells to a naïve state, in the presence of the t2iL+PKCi media formulation; t2iL+PKCi uses a titrated low dose 2i with LIF and a PKC inhibitor (Takashima et al., 2014). The t2iL+PKCi method has also been used to directly isolate preimplantation human naïve cells from the blastocyst ICM (Guo et al., 2016). The newly established culture conditions provide for the very first time an *in-vitro* model that enables us to expand our understanding of the earliest stages of human embryonic development.

Studies of human and mouse embryos have revealed that a global remodelling of the epigenome takes place during the transition between preimplantation and postimplantation development. Single and low cell number studies have mapped the transcriptional output of cells within the developing embryo and provide a trajectory upon which *in-vitro* states of pluripotency can be overlaid (Petropoulos et al., 2016; Stirparo et al., 2018). Human state specific expression of transcription factors such as KLF17 and DPPA5 in the naïve state and OTX2 and DUSP6 in the primed state are one of the hallmarks separating these two states of pluripotency (Collier and Rugg-Gunn, 2018). A prominent feature of preimplantation embryos is the global hypomethylated state, which is recapitulated upon reprogramming of primed hPSCs to the naïve state (Guo et al., 2014; Pastor et al., 2016). Conversely, primed hPSCs are globally hypermethylation, making methylation status an additional hallmark of naïve pluripotency (Collier and Rugg-Gunn, 2018). Further differences between naïve and primed hPSC have been noted in their metabolism, their X-chromosome status in female cells, and the distribution of H3K27me3 modification levels (Takashima et al., 2014; Theunissen et al., 2014, 2016; Vallot et al., 2017). Numerous studies have described the profound changes in transcription, protein expression, and the epigenetic landscape that take place during the short stage of development that is captured *in-vitro* using naïve and primed hPSCs as a model system. However, very little is known about the 3D

genome organisation of naïve human pluripotent cells and how *cis*-regulatory regions govern gene expression profiles with implications on human embryonic developmental progression.

5.1.1 Genome architecture and its role in cellular function

The enormous stretches of DNA within each cell's nucleus are not compacted in a random spaghetti fashion. Rather, early studies performed with electron microscopy revealed intricate structural organisation of DNA in metaphase chromosomes (Paulson and Laemmli, 1977). Subsequent research focusing on the interphase nucleus, with the use of fluorescent in-situ hybridisation, revealed that individual chromosomes occupy distinct nuclear territories and rarely intermingle (Cremer and Cremer, 2001; Cremer et al., 2006). Suggestions of the importance of genome architecture for function and transcriptional regulation came from early observations of the inactive X-chromosome and the formation of the Barr body (Barr and Bertram, 1977; Eils et al., 1996). However, the breakthrough that accelerated the study of genome architecture came with the development of the chromosome conformation capture (3C, one-to-one) technique (Dekker et al., 2002) (Figure 5-1). With formaldehyde fixation, restriction enzyme digestion and subsequent ligation of proximal fragments it became possible to investigate the interaction of linearly distant genomic loci at high resolution. Subsequent techniques derived from the principles of 3C expanded the ability of the technique to examine interactions of a specific viewpoint (4C, one-to-all) (Simonis et al., 2006; Zhao et al., 2006), a genomic region (5C, many-to-many) (Dostie et al., 2006) and ultimately of the entire genome (Hi-C, all-to-all) (Lieberman-Aiden et al., 2009) (Figure 5-1). The techniques, often complemented with single cell 3D FISH studies, allow the high-throughput discovery and quantification of interaction probabilities on a population and single cell level (Nagano et al., 2013).

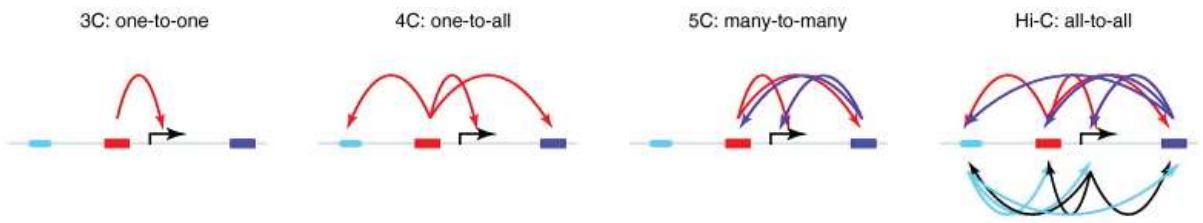


Figure 5-1: Chromosome conformation methods for the study of DNA interactions.
Black arrow designate the transcriptional start site. Arc arrows indicate the type and number of interaction each method can detect. The colour of the arrows matches the colour of the viewpoint rectangle. Taken from (Wijchers and de Laat, 2011).

The first genome wide conformation capture studies using Hi-C were also able to detect the distinct chromosomal territories that were previously described using chromosome paints (Lieberman-Aiden et al., 2009). In addition to chromosome territories, Hi-C was also able to detect two sets of genomic compartments within chromosomes at the 1-10 Mb scale, termed compartment A and B. The

underlying feature of these compartments was their preferential interaction with other regions of the same type, giving rise to a checkerboard pattern within interaction matrices. The compartments correlated with features such as DHS accessibility, gene density and histone marks, which led to the designation of the A compartment as the active, gene-rich euchromatin compartment, while the B compartment was an exact opposite with regions of inactive, gene-poor heterochromatin (Lieberman-Aiden et al., 2009). As large-scale structural features with a distinct chromatin status, the A/B compartments reflect the previously established tendencies of euchromatin and heterochromatin to segregate. These tendencies were first observed as late stage replication domains and more simply as the banding patterns of metaphase chromosomes (Allshire and Madhani, 2018). This perhaps suggests that the A/B compartments represent stable structural components of the genome. However, the capture of the dynamic genome reorganisation during hPSC differentiation revealed extensive compartment switching that was attributed to changes observed within topologically associated domains (TADs) (Dixon et al., 2015). With the increasing resolution and coverage of Hi-C studies, TADs were uncovered as yet another genome architectural feature at the lower sub-megabase level (Dixon et al., 2012; Hou et al., 2012; Nora et al., 2012; Sexton et al., 2012). The higher interaction frequency within a domain, with few interactions between domains is what defines TADs. The boundaries of TADs are enriched with cohesin, the protein complex that regulates the separation of sister chromatids during cell division, and the insulator protein CTCF (Dixon et al., 2012; Rao et al., 2014; Symmons et al., 2014), whose binding motif orientation was shown to determine its ability to form domains (Guo et al., 2015; Wit et al., 2015). The CTCF and cohesin dependent domain formation has been proposed to be the result of a loop extrusion mechanism, which is supported by perturbation experiments (Haarhuis et al., 2017; Wutz et al., 2017) and an *in-silico* model (Barrington et al., 2017; Fudenberg et al., 2016; Goloborodko et al., 2016). The large-scale principles of genome structure have provided us with enourmous insight into how DNA is folded within the nucleus and the implications thereof on the smaller scale contacts involved in transcriptional regulation.

The intricacy of transcriptional regulation is exemplified by the involvement of distal regulatory elements (enhancers, silencers, insulators) that have been shown to loop and contact non-adjacent promoters and ultimately impact their expression levels (Bulger and Groudine, 2011; Laat and Duboule, 2013; Spitz, 2016). Some of the most dramatic phenotypes include the sonic hedgehog (*Shh*) gene involved in limb development and the *Sox9* gene involved in sex determination. The *Shh* gene has a distal enhancer located 1 Mb away, in the intronic region of another gene. Point mutation or deletion of the *Shh* enhancer leads to the loss of *Shh* expression and results in limb deformation, which is similar in phenotype to the deletion of the *Shh* gene itself (Lettice et al., 2003; Sagai et al., 2005). In the *Sox9* example, the identification of a new enhancer Enh13 and its deletion substantially reduced

the levels of Sox9 expression in XY males below the required levels for testes development, leading to a sex reversal phenotype (Gonen et al., 2018). Additional studies with artificial tethering of enhancers and looping factors provided direct evidence of the importance of looping in the regulation of transcription (Deng et al., 2012; Müeller-Storm et al., 1989).

One of the limitations of Hi-C is the enormous diversity of pairwise interaction that it generates, requiring vast amount of sequencing to achieve enough coverage for the identification of interactions such as enhancers and promoters (Belton et al., 2012). To gain higher resolution information on the scale of promoter-enhancer interactions, a protein enrichment method based on Chromatin ImmunoPrecipitation (ChIP) was developed called Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET) (Fullwood et al., 2009). The method was successfully used to examine the interactome of active promoters bound with RNA polymerase II (Li et al., 2012). However, the inability to examine inactive promoters and regions of differential protein occupancy lead to the development of capture Hi-C (CHiC) and HiCap (Mifsud et al., 2015; Sahlén et al., 2015; Schoenfelder et al., 2015). The CHiC relies on a pool of capture probes consisting of 120 bp biotinylated oligos that enables a substantial enrichment of promoter interaction from a Hi-C library. The capture of promoter interactions allows the discovery of their distal regulatory elements and is revealing the structural organisation that leads to the coregulation of genes, such as through colocalization within transcription factories or repressive Polycomb bodies. Generation of extensive CHiC datasets in cells of the hematopoietic lineage additionally revealed that cell specific interactions can distinguish cell types, and can link non-coding disease associated variants with possible target genes (Javierre et al., 2016). Altogether, the studies of the 3D genomic architecture have revealed multiscale principles of DNA folding and its functional roles in controlling transcription.

5.2 Hypothesis

During the short developmental time between pre- and postimplantation development, cells within the human embryo undergo enormous epigenetic and chromatin remodelling. Some of these changes include the removal and reestablishment of DNA methylation marks, promoter histone modifications and the expression of a state specific transcriptome along with the translation of unique transcription factors. Based on these changes and other lower resolution studies performed in mESC (Joshi et al., 2015), it can be expected that there will be extensive promoter-enhancer rewiring.

5.3 Aims

Our current understanding of the 3D genome organisation of the human naïve and primed states of pluripotency is very limited. To better understand the coordination of developmental gene control, we have used promoter capture Hi-C to enrich over 22,000 promoters from naïve and primed hPSCs. My aims for this collaborative project are:

1. To examine new ways of visualising entire PCHiC datasets that would aid with an unbiased examination of interaction features.
2. To investigate the changes in promoter interactions between naïve and primed hPSCs.

5.4 A novel way of visualising and analysing promoter capture Hi-C (PCHiC) data

The enriched nature of PCHiC creates regions of locally enriched interactions that overall produce a relatively sparse significant interaction matrix. For this reason, PCHiC data is most commonly viewed and presented in the form of arc diagrams or circular plots, instead of classical HiC heatmaps. This limits the examination of the data to a select region of interest. For the general aim of uncovering interaction between a promoter and its distal regulatory elements such as enhancers, these methods are sufficient. However, a comprehensive way of examining entire networks of interacting promoters and other ends that could reveal co-regulatory mechanisms and elucidate the dynamic changes that separate different cell states has been lacking. Therefore, for the comparison of naïve and primed hPSCs promoter interactions I chose to combine both cell types into a single network, where the positions of HindIII fragments (nodes) would act as anchors around which the cell type specific interactions could be changed (Figure 5-2 a). Significant interactions only present within the naïve hPSCs are coloured blue, primed specific interactions are coloured red, while interactions present in both cell types (shared) are coloured black/grey. Using a force directed network layout, where nodes that interact more frequently are pulled closer together while less interacting nodes are pulled further apart, allowed me to visualise clusters of highly interacting HindIII fragments which tend to contain gene families such as protocadherin and the histone H1 genes (Figure 5-2 b). Importantly, it also allowed me to pick out clusters that preferentially interacted in naïve or in primed hPSCs. By examining the interaction frequency of each gene on a HindIII fragment, it became clear that the protocadherin genes have substantially more interactions in primed cells compared to naïve, while the opposite is true for the histone H1 genes (Figure 5-2 c). The increase in interactions for each family of genes also corresponded to an increased in the level of expression when compared to the other cell type (DESeq padj >= 0.05) (Figure 5-2 c and d). One of the powerful features of the network approach is that is allows me to overlay a range of different data onto the nodes (HindIII fragments) and edges (interactions between fragments). Asides from the origin (naïve or primed) of each interaction/HindIII fragment, shown in Figure 5-2 b, I have also overlaid categories of gene expression that again highlights the protocadherin and histone H1 clusters as predominantly expressed in naïve or in primed (Figure 5-2 d). The expression overlay also allows quick identification of other biologically relevant clusters and eliminate perhaps less functionally important clusters such as the keratin and olfactory receptor clusters, which despite having differences in interaction frequencies have no differential

gene expression. An overlay of A/B compartment switching reveals clusters of genes that undergo activation (B to A) or silencing (A to B) (Figure 5-3). One prominent example is the *TET2* gene that encodes the enzyme responsible for converting DNA 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC), which form part of the active DNA demethylation cycle (Wu and Zhang, 2017). *TET2* is within the active A compartment in naïve cells and switches to the inactive B compartment in primed cells. The expression of the TET2 enzyme could partly explain the DNA hypomethylation observed in naïve hPSCs (Collier and Rugg-Gunn, 2018). Altogether, the ability to visualise entire PCHiC datasets and examine clusters of interactions provides substantial advantages over previous methods and has allowed us to examine previously unappreciated clusters of promoter interactions.

The PCHiC datasets also reveal previously undocumented interactions with regulatory regions in the naïve state that correlate with higher expression of state-specific genes. Two examples include the *TET2* and *DPPA5* genes. The HindIII fragment containing the *TET2* promoter in naïve hPSCs forms several interactions with downstream enhancer regions that are enriched in H3K27ac (Figure 5-4 – highlighted in grey). These interactions correlate with higher *TET2* gene expression naïve compared to primed hPSCs. Moreover, the slight increase of repressive H3K27me3 over the *TET2* promoter region in primed hPSCs may contribute to the lower levels of transcription in primed cells. The DPPA5 gene is one of the highest differentially expressed genes between naïve and primed hPSCs (Collier et al., 2017; Stirparo et al., 2018). In naïve hPSCs, the HindIII fragment containing the *DPPA5* promoter establishes naïve-specific interactions with several downstream naïve-specific enhancer regions enriched with H3K27ac (Figure 5-5 – highlighted in grey). Besides these, the entire region surrounding *DPPA5* in naïve hPSCs is in an active state with an increased number of interactions. These examples demonstrate the value of the PCHiC data as a resource for better understanding the naïve state of pluripotency. Further exploration of this data will help to reveal the *cis*-regulatory regions that physically contact and influence transcription of developmentally important genes.

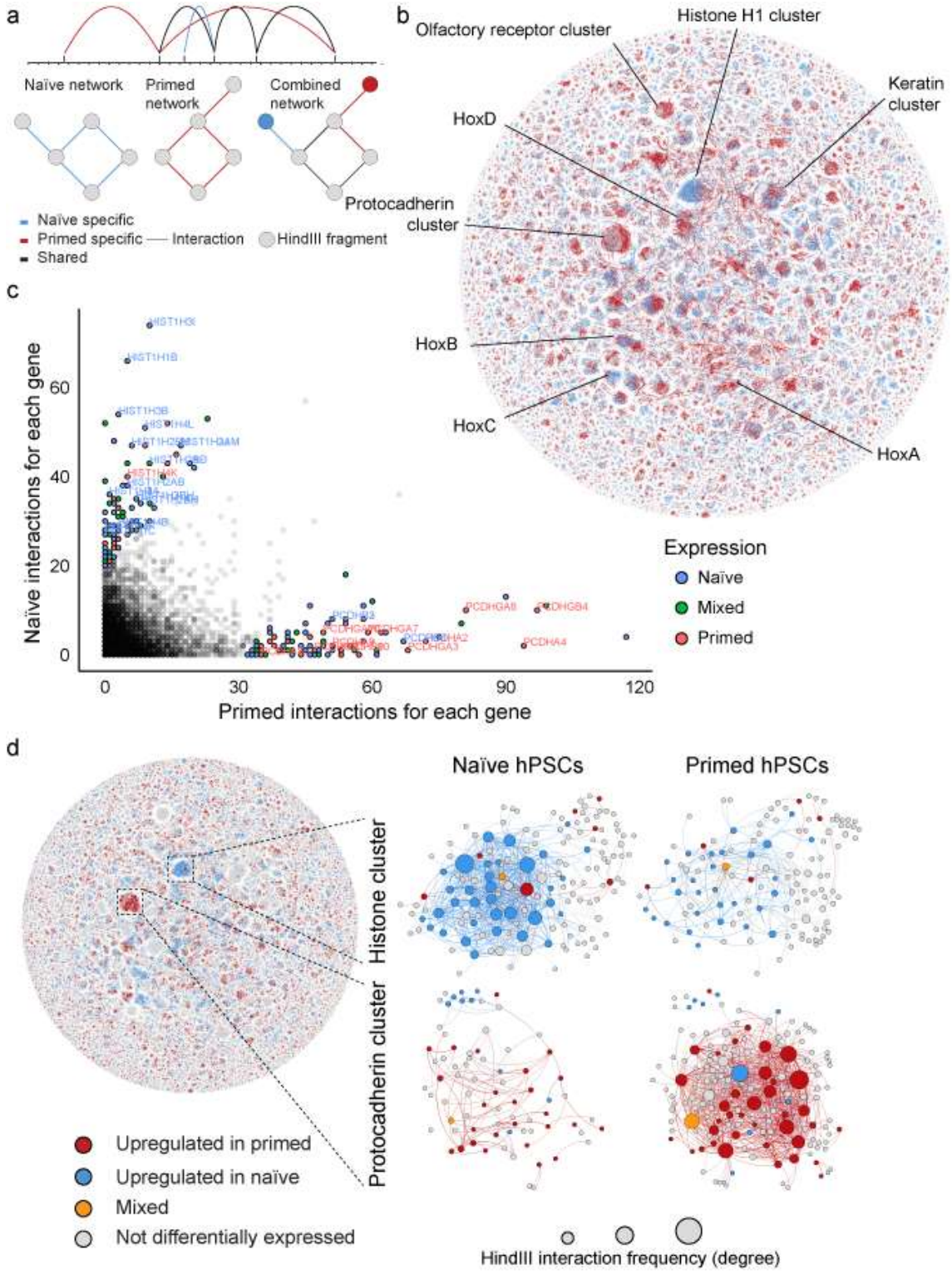


Figure 5-2: Network visualisation of PCHiC data reveals large-scale interaction changes between two samples.
(a) Arc diagrams are the typical PCHiC visualisation method, which however, only allow the examination of a limited region of the genome. By constructing networks, where each node represents a HindIII fragment and each edge represents a significant interaction called by CHiCAGO (see Methods, section 2.16.2), it is possible to get an overview of entire PCHiC datasets. A singular network is created from both datasets where interactions unique to naïve hPSC are drawn in blue, primed hPSC are drawn in red, and interactions shared by both are drawn in black/grey. The same colour scheme is also applied to HindIII fragments. (b) A force directed layout (Jacomy et al., 2014) of the network containing both the naïve and primed samples. Some of the interaction clusters containing gene clusters have been labelled. (c) The number of interaction that each gene establishes with other HindIII fragments highlights some of the most highly interacting genes. (d) Differential gene expression categories overlaid onto the network. The size of the nodes corresponds to their interaction frequency.

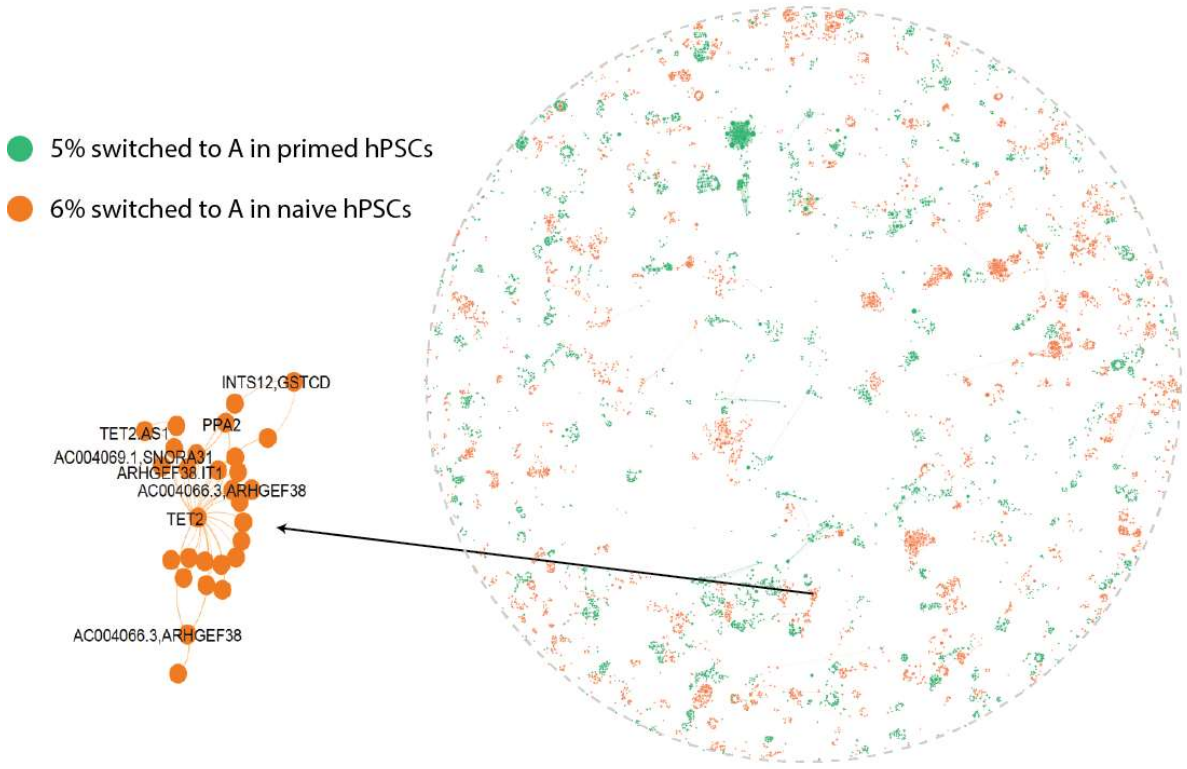


Figure 5-3: Switching from active to inactive compartments happens in several interaction hubs that include the TET2 gene. Approximately 11% of HindIII fragments switch compartments between the naïve and the primed cell states. 5% of compartment B (inactive) fragments in naïve cells switch to compartment A (active) in primed cells (green), while 6% of compartment A fragments in naïve cells switch to compartment B in primed cells (orange).

5.4.1 Force directed layout captures general features of genome folding

Significant PCHiC interactions are centred on baited fragments that produce island of enrichment. Due to these local *cis* interactions the PCHiC network is not fully connected and instead contains numerous independent subnetworks. These subnetworks in turn contain clusters of highly interacting bait and other ends that can be subdivided into what is known in network science as communities (Figure 5-6). The communities form interactions within themselves more frequently then with other communities. This is essentially the definition of topologically associated domains (TADs) (Dixon et al., 2012), which led me to compare TAD calls from our HiC data with the communities detected in the PCHiC network. Overlaying TAD information, determined from naïve and primed HiC, onto the network nodes reveals a striking concordance with subnetwork and community clusters (data not shown). I subsequently examined the overlap of HindIII fragments within communities and TADs. The TAD-community overlap was significantly higher than compared to random regions of equal TAD size (Figure 5-7). Altogether, the comparison of TADs and communities demonstrates that the clusters created with the force-directed layout capture general features of genome folding.

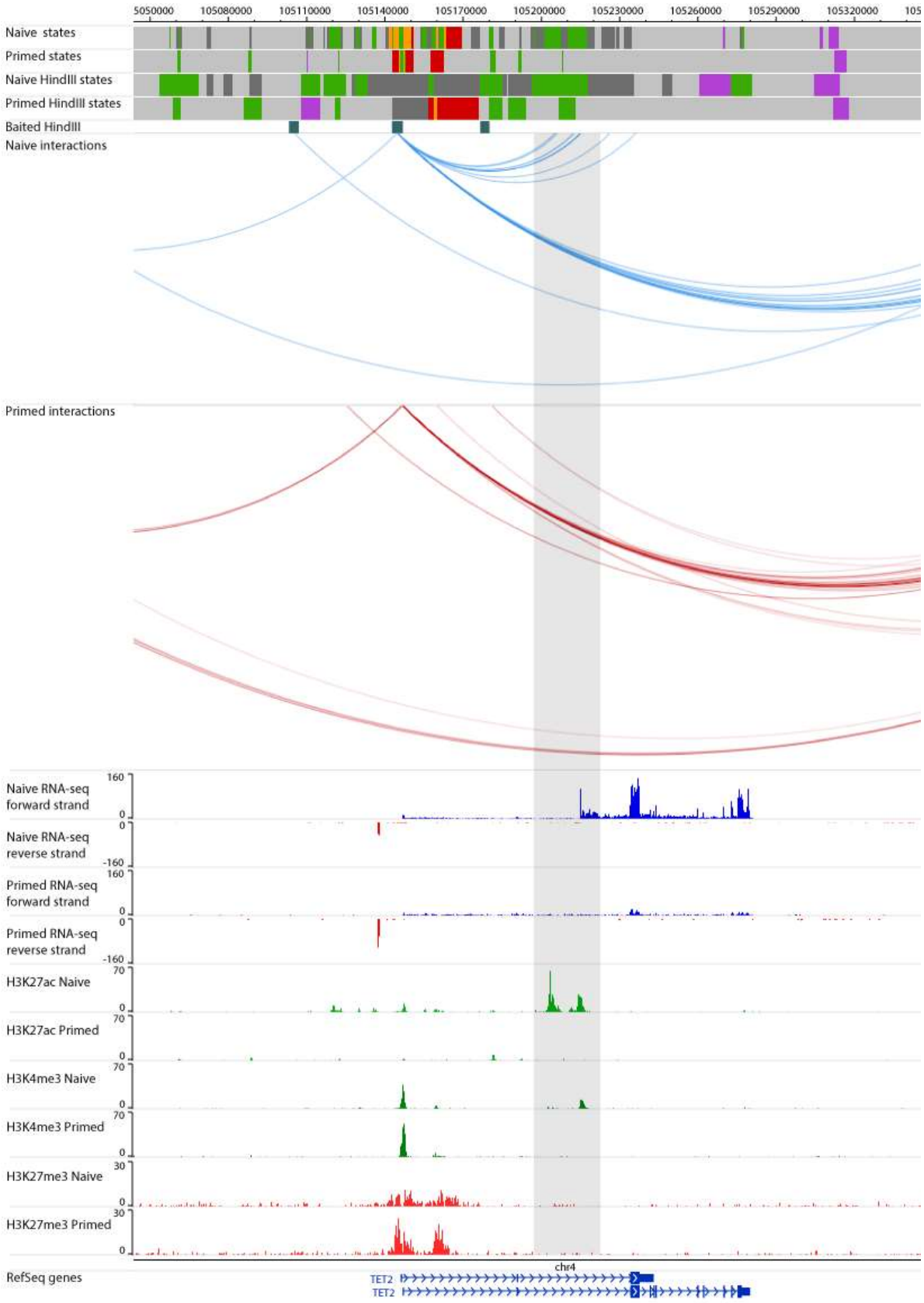


Figure 5-4: TET2 establishes unique interactions to distal regulatory elements in the naïve state.
Grey rectangle highlights naïve specific interaction to distal regulatory elements. Tracks were visualised using the WashU epigenome browser. The values of RNA- and ChIP-seq represent normalised reads counts (see Methods, section 2.16.6). Chromatin states identified by ChromHMM are coloured as follows: active – green; bivalent – orange; heterochromatin – purple; Polycomb-only - red; mixed – yellow; background – grey; unclassified – dark grey.

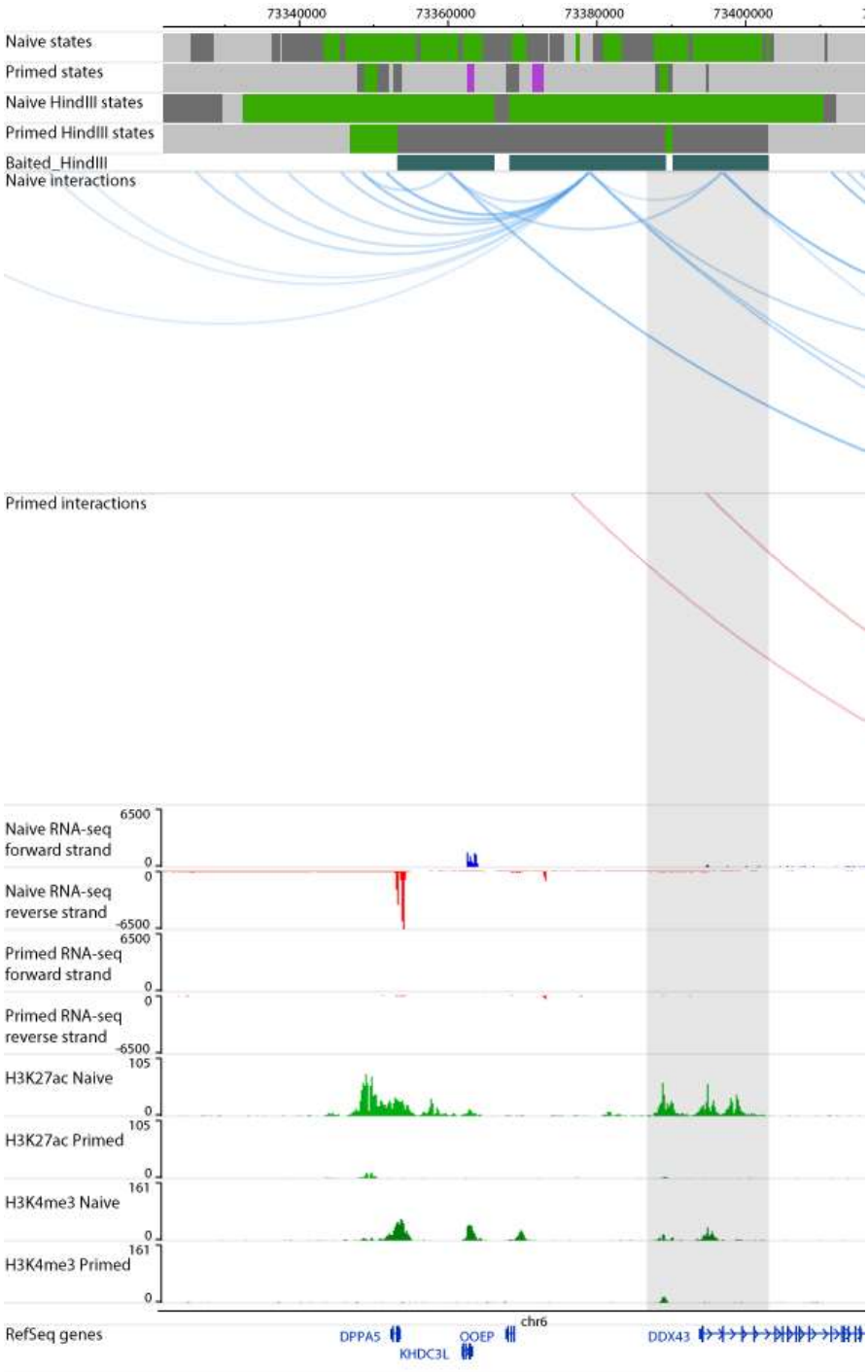


Figure 5-5: DPPA5 establishes unique interactions to distal regulatory elements in the naïve state.
Grey rectangle highlights naïve specific interaction to distal regulatory elements. Tracks were visualised using the WashU epigenome browser. The values of RNA-seq and ChIP-seq represent normalised reads counts (see Methods, section 2.16.6). Chromatin states identified by ChromHMM are coloured as follows: active – green; bivalent – orange; heterochromatin – purple; Polycomb-only - red; mixed – yellow; background – grey; unclassified – dark grey.

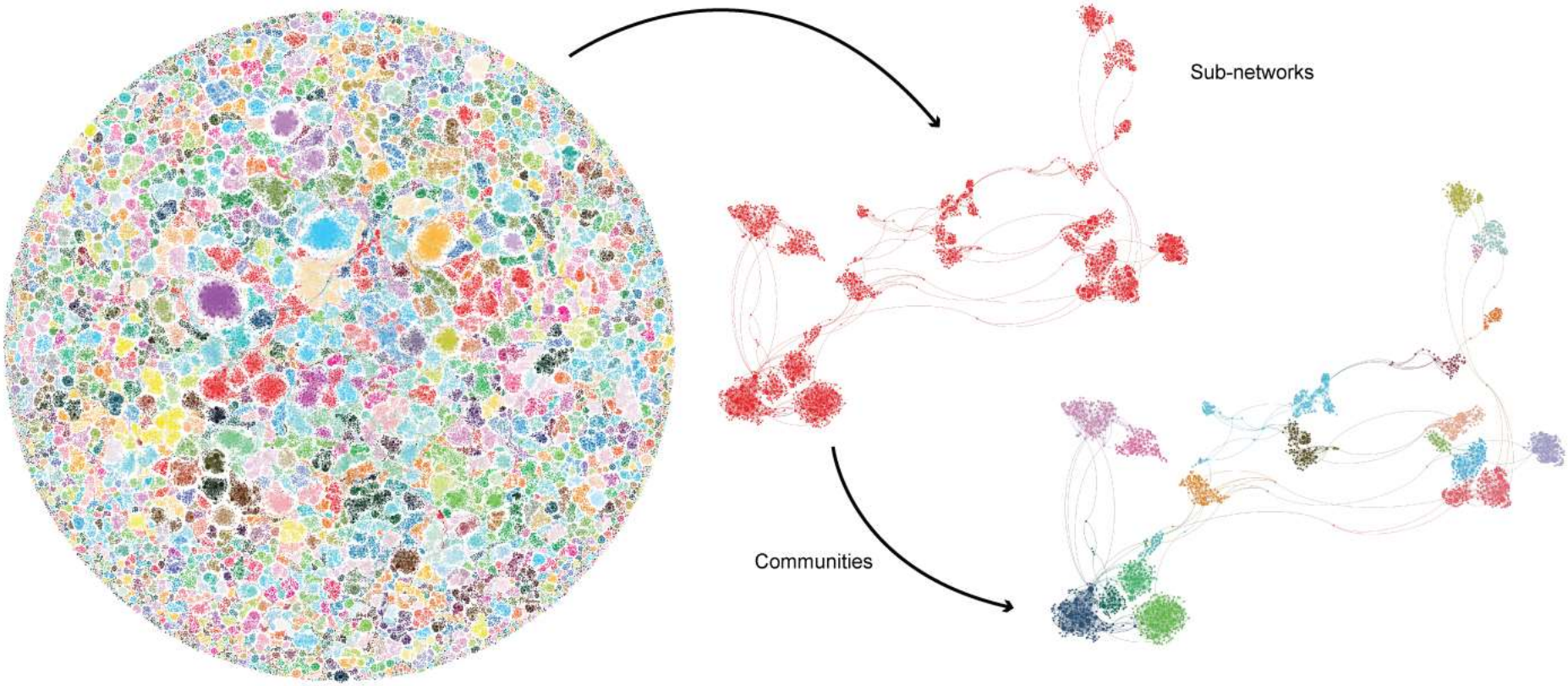


Figure 5-6: The PCHiC network can be divided into subnetworks and individual communities.
Each subnetwork is highlighted with a unique colour. Large subnetworks can be further broken down into communities. The multiple levels of the PCHiC network can be leveraged to perform comparative analysis between naïve and primed hPSC and gain insight into which clusters of interacting genes change in structure and composition.

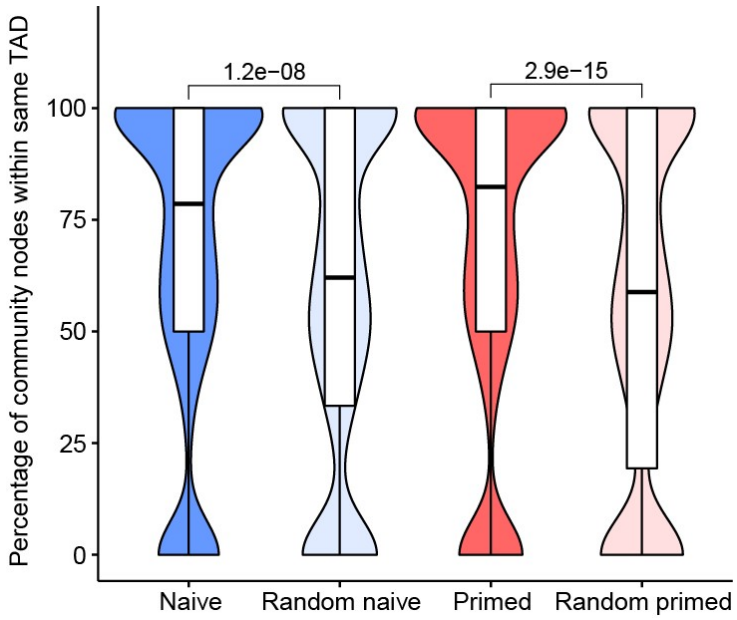


Figure 5-7: Network communities overlap with TADs to a significantly greater extent than expected by chance. The percentage overlap between HindIII fragments within communities and the HindIII fragments within TADs was compared to the overlap with randomly shuffled regions matching the size of each TAD. Unpaired two-sample Mann-Whitney/Wilcoxon rank sum test with continuity correction was used to test the significance of the overlap difference.

5.5 Examination of subnetwork dynamics reveals long range interactions as a major contributor to the differences between naïve and primed hPSCs

One of my main aims in merging the naïve and primed PCHiC data into a single network was to examine interaction differences. Initially, I was more interested in large interaction changes in groups of genes that would suggest co-regulation, rather than subtle rewiring of enhancer regions that may influence transcription of individual genes. I chose to gauge the dynamics between naïve and primed hPSCs at the subnetwork level and decided to use the difference in the number of cell type specific interactions and HindIII fragments within each subnetwork as a measure of change (Figure 5-8 a); the greater the distance from the origin, the greater the change between the two cell states. Reassuringly, the protocadherin and the histone H1 subnetworks once again stood out. Interestingly, the size of numerous primed subnetworks was greater compared to their counterparts in naïve hPSCs, as highlighted by the skew of subnetworks in the lower-left quadrant of Figure 5-8 a. The skew suggests that primed cells establish interactions that bring together multiple communities that are self-contained in naïve cells. To investigate the cause of the community merging into a larger subnetwork in primed hPSCs, I first examined the most changing subnetwork (Figure 5-8 a). Linear genomic plots and a multidimensional scaling network layout revealed that primed cells establish long-range interactions (> 1 mb) that are absent in naïve cells (Figure 5-8 b). I next wanted to determine whether the establishment of long-range interactions was unique to the examined subnetwork, or if it was a more widespread feature. Examining the mid-range and shortest 1000 interactions in each dataset revealed no appreciable differences between naïve and primed. However, the top 1000 longest

interactions were considerably longer in primed cells compared to naïve, suggesting this is a general feature (Figure 5-8 c).

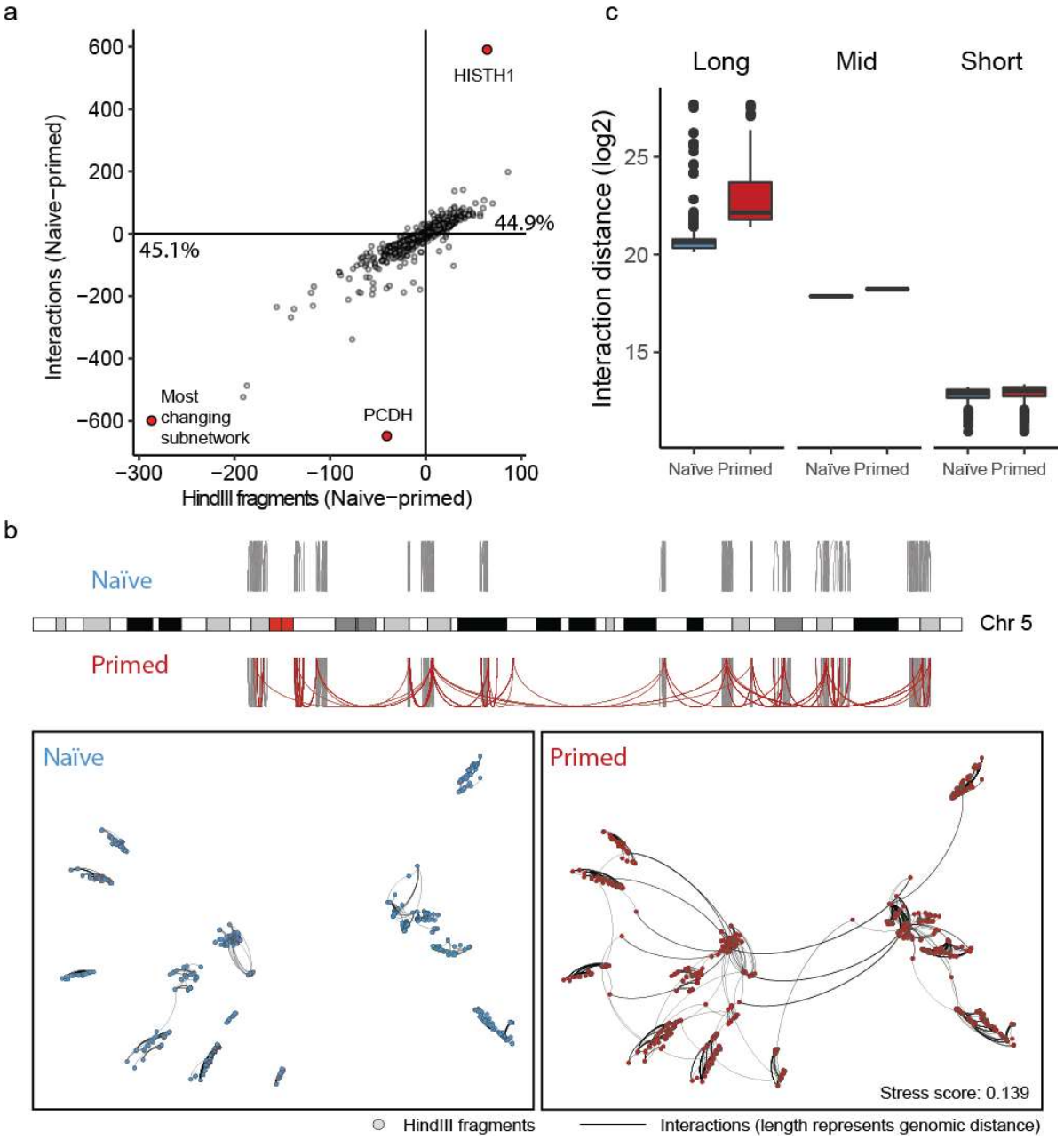


Figure 5-8: Large-scale changes between naïve and primed hPSCs result from the gain of long-range interaction in primed hPSCs.
(a) Difference in the number of interactions and interacting HindIII fragments between naïve and primed hPSC reveals the most changing subnetworks. The lower-left quadrant contains subnetworks with more interactions and interacting HindIII fragments in primed hPSC, while the top-right quadrant contains larger naïve hPSC subnetworks. The protocadherin (PCDH) and histone H1 (HISTH1) subnetworks are highlighted in red along with the most changing subnetwork. (b) The most changing network plotted on a linear genomic arc plot (top) with interaction longer than 2^{20} (> 1 mb) highlighted in red in primed and blue in naïve (no blue interactions are present). By using multi-dimensional scaling (MDS) with the genomic distance as the edge weights, the distance between nodes can be interpreted as linear genomic distance. The stress score denotes the reliability or degree of correspondence of the genomic distances and the distances between the nodes determined by MDS (score of 0-1). Values below 0.1 represent a good fit, whereas values above 0.15-0.2 are generally too far from the original distances to make meaningful conclusions (Kruskal and Wish, 1978). (c) Comparison of 1000 longest, shortest and mid-range interaction distances between naïve and primed hPSC (long - > 1mb; mid – around 15 kb; short – around 6 kb).

To establish what the long-range interactions are doing and whether they are biologically relevant, I used ChromHMM (Ernst and Kellis, 2012, 2017) to determine the chromatin state of HindIII fragments with the top 1000 longest interactions in naïve and in primed hPSCs (Figure 5-9 a). In naïve cells, most long-range interactions were between active state (H3K4me3, H3K27ac, H3K4me1) fragments. Conversely, in primed cells the longest interactions formed between bivalent state (H3K27me3, H3K4me3, H3K4me1) fragments. With the dominance of the bivalent state in primed cells, it is possible that the Polycomb group (PcG) proteins, namely the Polycomb repressive complex 2 (PRC2) responsible for depositing H3K27me3 marks via the EZH1/2 methyltransferase subunit (Schuettengruber et al., 2017), could be driving the establishment of the observed long-range interactions. To further examine the regions of PRC2 occupancy, I overlaid the bivalent and Polycomb chromatin states onto separate naïve or primed interaction networks (Figure 5-9 b). In primed hPSCs, the two chromatin states overlap with several HindIII fragment clusters that contain numerous interactions within and between each other. Reassuringly, these interaction hubs contain homeobox genes, including the HOXA-D gene clusters, which encode a set of transcription factors responsible for specifying segment identity within the embryo; in other words, specifying which part of the embryo will give rise to which body part (Pearson et al., 2005). Homeobox genes have been previously shown to be regulated by PcG and Trithorax group proteins (Schuettengruber et al., 2017). Based on the large concentration of co-regulated HindIII fragments, it is tempting to speculate that the hubs correspond to Polycomb bodies, which have been described by microscopy as nuclear foci with high concentrations of PcG proteins (Matheson and Elderkin, 2018). However, further work will need to be performed to confirm this hypothesis. Examining the same regions in naïve cells reveals a stark contrast to primed cells, with most of the hub and inter-hub interactions being absent (Figure 5-9 b). This is perhaps not surprising given the absence of the H3K27me3 mark from the promoters of naïve cells, which has been established in a number of studies in both human and mouse (Marks et al., 2012; Reddington et al., 2013; Theunissen et al., 2014). Altogether, the absence of H3K27me3 repressive hubs in naïve cells contrasts with the widespread, long-range, highly interacting hubs observed in primed cells.

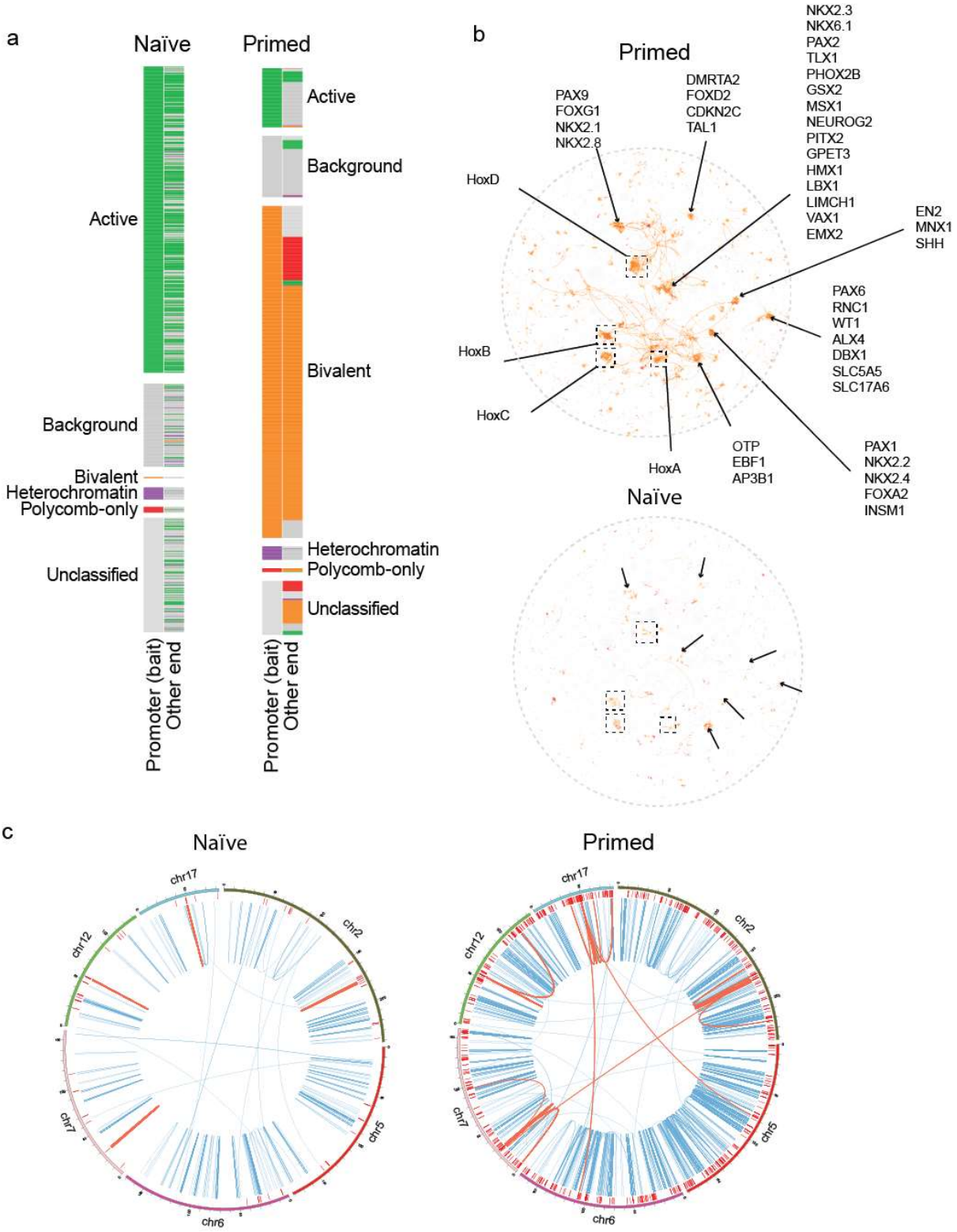


Figure 5-9: Comparison of chromatin states of top 1000 longest interactions reveals dominance of the bivalent state in primed hPSCs.

Chromatin states were determined using ChromHMM. (a) The chromatin states of the nodes connected by the top 1000 longest interactions in naïve and the top 1000 in primed. Each heatmap is divided by the chromatin state of the promoter HindIII fragment. (b) Visualising bivalent and Polycomb-only interactions on the network highlights several clusters in primed hPSC (top) that all contain homeobox genes. The clusters, together with the interactions are virtually absent in naïve hPSC (bottom). (c) Circos plots of the 6 chromosomes containing HOX genes. Interactions with and between HOX genes are highlighted in red. Other long range (2^{20}) and trans interactions are shown in blue. Track on the outside shows H3K27me3 peaks. HOXA – chr7; HOXB – chr17; HOXC – chr12; HOXD – chr2.

5.6 Intra-*HOX* gene interaction are absent from naïve hPSCs

The stark contrast of interactions between naïve and primed hPSCs in the HOX clusters led me to examine them in more detail. Circos plots reveal, in more detail, the almost complete absence of interactions in naïve cells, except for directly above the HOX clusters, while primed cells display numerous long-range and *trans* interactions (Figure 5-9 c). Similar observations were also made in the equivalent naïve and primed mouse system (Joshi et al., 2015). However, my closer examination of the HOXA cluster revealed a surprising absence of intra-HOXA interactions in naïve hPSCs, which were a prominent feature in primed hPSCs (Figure 5-10). The change in intra-HOXA interactions was accompanied by a decrease of the H3K27me3 mark, while the H3K4me3 remained equivalent in both cell types. The decrease in H3K27me3 resulted in active chromatin state annotation of small gene overlapping regions, along with an observable leaky transcription of *HOXA* genes. I further examined the HOXC cluster, where I observed a similar absence of intra-HOXC interactions (Figure 5-11 a). However, the inter-HOXC interactions appear to be more prominent in naïve compared to primed cells. Quantification of the inter-HOX and intra-HOX interactions revealed that HOXA and HOXD have more inter-HOX interactions in primed cells, while HOXB and HOXC have more inter-HOX interactions in naïve cells (Figure 5-11 b). Overall, intra-HOX interactions are a dominant feature of primed hPSCs, while inter-HOX interactions are observed more equally between the two cell types, indicating that intra-HOX interactions may be established as a means to prime hPSCs for subsequent differentiation during development.

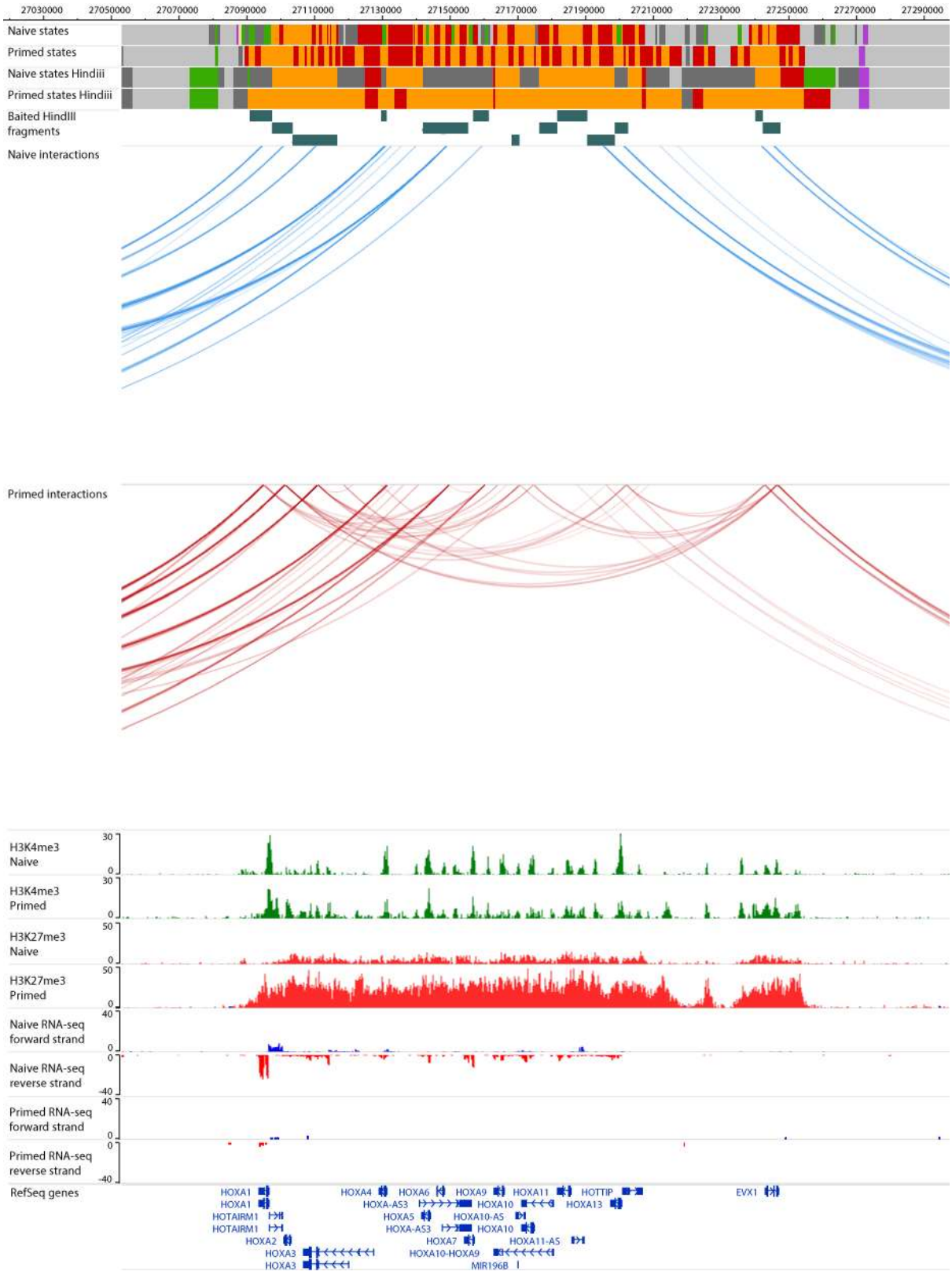


Figure 5-10: Visualisation of HOXA interactions and histone modifications reveals differences between naïve and primed hPSCs.

Interactions with (inter) and between (intra) HOXA genes reveals a dominance of intra cluster interactions in primed hPSCs. Tracks were visualised using the WashU epigenome browser. The values of RNA- and ChIP-seq represent normalised reads counts (see Methods, section 2.16.6). Chromatin states identified by ChromHMM are coloured as follows: active – green; bivalent – orange; heterochromatin – purple; Polycomb-only - red; mixed – yellow; background – grey; unclassified – dark grey.

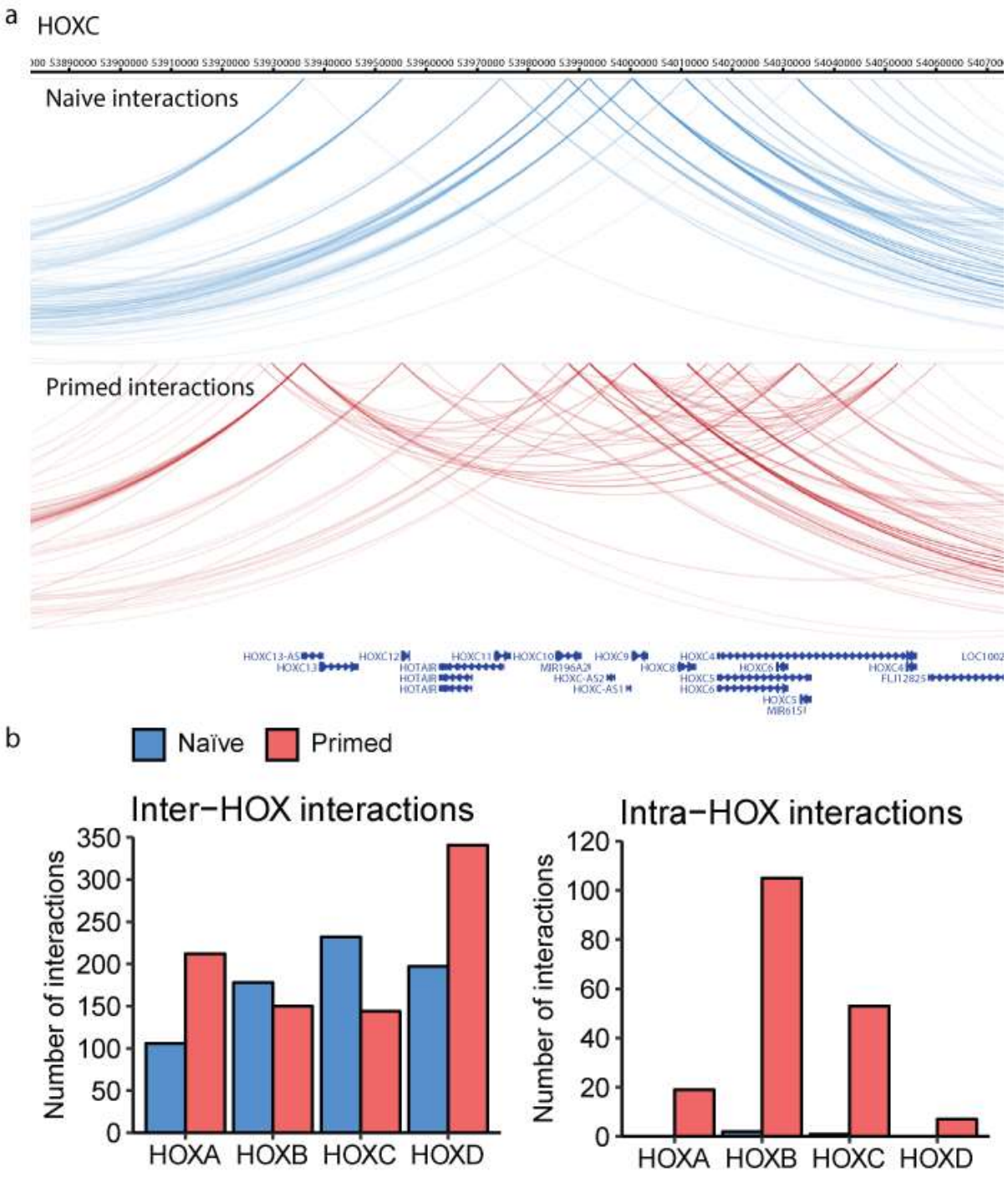


Figure 5-11: Intra-HOX interactions dominate in primed hPSCs.
(a) Comparison of interactions between naïve (blue) and primed (red) hPSCs in the HOXC gene cluster. Bottom track shows RefSeq genes. (b) Quantification of the number of interactions that HOX genes make with other non-HOX containing (inter) and HOX containing fragment (intra).

5.7 Discussion

The examination of promoter interactions in naïve and primed hPSCs using a network approach revealed numerous, large connected hubs with cell state-specific interaction frequency. The use of a force directed network layout (Jacomy et al., 2014) revealed features of genome organisation, such as TADs. This is perhaps not surprising given that several network-based approaches have been recently proposed as an accurate means for calling TADs (Norton et al., 2018; Yan et al., 2017). Nonetheless, the ability to view entire PCHiC datasets with meaningful organisational features on multiple scales provides an enormously powerful discovery tool. Indeed, by quantifying the differences in interacting HindIII fragments and total interactions on the scale of sub-networks, it was possible to identify several sub-networks between naïve and primed hPSCs that underwent substantial remodelling. Additionally, analysis of the changes in connectivity between individual communities within sub-networks also revealed a higher propensity of small communities to come together and form much larger interacting units in primed hPSCs. By overlaying histone-modification datasets onto HindIII fragments, it was apparent that long-range *cis*-interactions marked with H3K27me3 are associated with the aggregation of large interaction hubs. Together with future planned immuno-FISH experiments, these findings could establish a link with Polycomb body formation in the establishment of these large interacting units. Overall, the network-based approach to visualising and analysing PCHiC datasets has provided us with a multi-scale overview, ranging from global interaction dynamics to local communities that recapitulate TADs, and all the way down to individual promoter-enhancer interactions.

Some of the sub-networks that underwent substantial remodelling include the histone H1 and the protocadherin gene families. Variants of the histone H1 family have previously been shown to be important in the maintenance of pluripotency, the ability to differentiate and in the repression of satellite repeats in mESCs and human induced pluripotent stem cells (iPSCs) (Turinetto and Giachino, 2015). The hypomethylated state of naïve hPSCs would be expected to lead to de-repression of repeats, suggesting that higher expression of the histone H1 genes in naïve hPSCs, possibly resulting from their increased interactivity, could be a compensatory repression mechanism for satellite repeats. On the other hand, the protocadherin gene family encodes cell adhesion molecules that are predominantly expressed in the neural lineage (Chen and Maniatis, 2013; Sano et al., 1993). Based on DamID studies, the protocadherin cluster is associated with the nuclear lamina (a protein mesh structure on the inner nuclear membrane) (Kind et al., 2015), which could account for the relatively low expression and interaction in naïve hPSCs. In primed hPSCs, the increased expression of several protocadherin genes could entail their selective looping out of the repressive lamina environment, which based on their increased flexibility would increase their interactivity with other genes within

the protocadherin cluster. Altogether, the interaction dynamics of sub-networks reveal new associations with the switching of pluripotent states.

In addition to some of the large-scale interaction differences observed between naïve and primed hPSC, the PCHiC datasets have also revealed numerous examples of enhancer switching. In mouse naïve and primed ESCs the switching of enhancers has been previously documented and used as marker for distinguishing the two cell states (Collier and Rugg-Gunn, 2018). The OCT4 gene in naïve mESCs uses the distal enhancer, while in primed mEpiSCs the proximal enhancer is utilised instead. The use of a 6-cutter such as HindIII leads to lower resolution compared to a 4-cutter such as MboI (Sahlén et al., 2015). Because of this, proximal enhancer interaction in our datasets will likely not be detected. Nonetheless, the advantage of the less frequent cutter is its ability to capture longer-range interactions. Some of the genes that changed their distal enhancers between naïve and primed hPSCs includes *TET2* and the naïve specific gene *DPPA5*. The current analysis of PCHiC with the CHiCAGO pipeline is based on pairwise comparisons of all the fragments (Cairns et al., 2016), which in some cases can produce stretches of significant interaction fragments, from which it can be challenging to infer the most important interactions. To address this complication, a method was developed that uses fine mapping, typically employed in genome-wide association studies (GWAS), to identify the primary interacting fragment (Eijsbouts et al., 2018). The method was shown to improve the resolution of existing datasets. The data presented here provide the opportunity for follow-up enhancer knockout studies to functionally interrogate the role of altered enhancer interactions, and fine-mapping could potentially increase the success of such follow-up studies. Altogether, our PCHiC datasets and analysis provide a useful resource for understanding how cells are primed into lineage commitment and make one of their first developmental decisions.

The observed emergence of intra-HOX interactions in primed hPSCs correlates with PRC2 H3K27me3 modification, which together with H3K4me3 establishes a poised environment in which HOX genes are silent until ready for expression in a temporally controlled manner. The highly self-interacting HOX domains are one of the best examples of TAD regulated gene expression (Andrey et al., 2013), and their compaction into large hubs has been shown to involve interactions beyond simply pairwise that are detected by classical 3C methods (Olivares-Chauvet et al., 2016). These previous observations appear to be captured with our network approach, where the expected regulation by repression results in highly interacting hubs that include the HOX loci. The expression of Polycomb group proteins and the respective demethylases are similar between the two states, while the global H3K27me3 levels remain the same (Marks et al., 2012). This observation has also been made in human naïve and primed pluripotent stem cells (Theunissen et al., 2014). However, the absence of H3K27me3 at promoters of key developmental genes in naïve hPSCs raises the question of what is keeping them

silent and preventing cells from improperly differentiating. Other studies have shown the involvement of PRC1 in the formation of long-range interaction involving the HOX genes (Schoenfelder et al., 2015). This requirement for PRC1 in long-range interactions, the loss of H3K27me3 at promoters of naïve hPSC and the absence of promoter long-range interactions in our PCHiC data, suggests that an interplay between both Polycomb repressive complexes may be required to establish highly interacting repressive hubs. A previous study in naïve and primed mESCs has also demonstrated the importance of PRC2 in the establishment of long range interactions and as the driver of genome organisational differences between the two cell states (Joshi et al., 2015). Further experiments examining the site of PRC1 occupancy and potentially other repressive complexes, such as REST, could potentially reveal the mechanism by which these important developmental loci are silenced in naïve hPSCs and which factors are directly involved in the formation of long-range interactions. A more recent study, using CRISPR-Cas9, demonstrated a PRC2 dependence in the establishment of poised enhancer interaction with their target genes (Cruz-Molina et al., 2017). The absence of PRC2 was shown to compromise the induction of genes in the differentiated cell types, suggesting a function of PRC2 beyond just repression that includes genome organisation and proper activation of genes required in differentiation.

Summary of finding:

- Network visualisation of PCHiC data revealed large-scale changes in interaction hubs between naïve and primed hPSCs.
- Rewiring of promoter-enhancer interactions of cell specific genes takes place and correlates with transcriptional changes, with *TET2* and *DPPA5* as examples.
- Large-scale changes between naïve and primed hPSCs stem from the gain of long-range interactions in primed hPSCs, which are dominantly established between HindIII fragments with the bivalent and Polycomb chromatin states.
- Large interacting hubs formed by long range interactions in primed hPSCs overlap with homeobox genes, including the four HOX clusters that display intra-HOX interactions in primed hPSCs, which are absent in naïve hPSCs.

Large amounts of starting material have been a limiting factor for examining the genome architecture of rare populations such as pro-B cells. Recent improvements in the Hi-C protocol are now enabling the study of chromatin conformation in thousands of cells instead of the previously required tens of millions of cells (Blanco et al., 2018; Oudelaar et al., 2017). We have recently published PCHiC datasets from pre-B cells, which examined the changed in genome architecture that take place during ageing (Koohy et al., 2018). A number of differences were observed that included

the A to B compartment switching of the genomic region encompassed the *Irs1* gene. Future efforts are now focused on performing an equivalent capture in pro-B cells. The network approach of visualising the entire PCHiC dataset will provide a valuable tool for data exploration and analysis, comparing pro-B and pre-B datasets. The global overview of combined datasets could enable the identification of differences at individual loci, which may have been missed otherwise. The future use of dynamic networks could additionally provide a new insightful way of visualising the changes in genome architecture as B cells transition through the different developmental stages.

6 Bibliography

Afshar, R., Pierce, S., Bolland, D.J., Corcoran, A., and Oltz, E.M. (2006). Regulation of IgH Gene Assembly: Role of the Intronic Enhancer and 5'DQ52 Region in Targeting DHJH Recombination. J. Immunol. *176*, 2439–2447.

Agrawal, A., and Schatz, D.G. (1997). RAG1 and RAG2 Form a Stable Postcleavage Synaptic Complex with DNA Containing Signal Ends in V(D)J Recombination. Cell *89*, 43–53.

Alamyar, E., Duroux, P., Lefranc, M.-P., and Giudicelli, V. (2012). IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. Methods Mol. Biol. Clifton NJ *882*, 569–604.

Alamyar, E., Giudicelli, V., Duroux, P., and Lefranc, M.-P. (2014). Antibody V and C domain sequence, structure, and interaction analysis with special reference to IMGT®. Methods Mol. Biol. Clifton NJ *1131*, 337–381.

Alanen, A., Pira, U., Lassila, O., Roth, J., and Franklin, R.M. (1985). Mott cells are plasma cells defective in immunoglobulin secretion. Eur. J. Immunol. *15*, 235–242.

Allman, D., and Pillai, S. (2008). Peripheral B cell subsets. Curr. Opin. Immunol. *20*, 149–157.

Allman, D., Lindsley, R., W, D., Rudd, K., Shinton, S., and Hardy, R. (2001). Resolution of three nonproliferative immature splenic B cell subsets reveals multiple selection points during peripheral B cell maturation. J. Immunol. Baltim. Md 1950 *167*, 6834–6840.

Allman, D.M., Ferguson, S.E., Lentz, V.M., and Cancro, M.P. (1993). Peripheral B cell maturation. II. Heat-stable antigen(hi) splenic B cells are an immature developmental intermediate in the production of long-lived marrow-derived B cells. J. Immunol. *151*, 4431–4444.

Allshire, R.C., and Madhani, H.D. (2018). Ten principles of heterochromatin formation and function. Nat. Rev. Mol. Cell Biol. *19*, 229–244.

Alt, F.W., Yancopoulos, G.D., Blackwell, T.K., Wood, C., Thomas, E., Boss, M., Coffman, R., Rosenberg, N., Tonegawa, S., and Baltimore, D. (1984). Ordered rearrangement of immunoglobulin heavy chain variable region segments. EMBO J. *3*, 1209–1219.

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs. Science *340*, 1234167.

Arnold, L., Pennell, C., McCray, S., and Clarke, S. (1994). Development of B-1 cells: segregation of phosphatidyl choline-specific B cells to the B-1 population occurs after immunoglobulin gene expression. J. Exp. Med. *179*, 1585–95.

Baker, N., and Ehrenstein, M.R. (2002). Cutting edge: selection of B lymphocyte subsets is regulated by natural IgM. J. Immunol. Baltim. Md 1950 *169*, 6686–90.

Bannas, P., Hambach, J., and Koch-Nolte, F. (2017). Nanobodies and Nanobody-Based Human Heavy Chain Antibodies As Antitumor Therapeutics. Front. Immunol. *8*.

Barbas, S.M., Ditzel, H.J., Salonen, E.M., Yang, W.P., Silverman, G.J., and Burton, D.R. (1995). Human autoantibody recognition of DNA. Proc. Natl. Acad. Sci. U. S. A. *92*, 2529–2533.

Barr, M.L., and Bertram, E.G. (1977). A Morphological Distinction between Neurones of the Male and Female, and the Behaviour of the Nucleolar Satellite during Accelerated Nucleoprotein Synthesis. In Problems of Birth Defects: From Hippocrates to Thalidomide and After, T.V.N. Persaud, ed. (Dordrecht: Springer Netherlands), pp. 101–102.

Barrington, C., Finn, R., and Hadjur, S. (2017). Cohesin biology meets the loop extrusion model. Chromosome Res. *25*, 51–60.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks.

Baum, P.D., Venturi, V., and Price, D.A. (2012). Wrestling with the repertoire: The promise and perils of next generation sequencing for antigen receptors. Eur. J. Immunol. *42*, 2834–2839.

Baumgarth, N. (2011). The double life of a B-1 cell: self-reactivity selects for protective effector functions. Nat. Rev. Immunol. *11*, 34–46.

Beisel, C., and Paro, R. (2011). Silencing chromatin: comparing modes and mechanisms. Nat. Rev. Genet. *12*, 123–35.

Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi–C: A comprehensive technique to capture the conformation of genomes. Methods *58*, 268–276.

Bendall, S.C., Simonds, E.F., Qiu, P., Amir, E.D., Krutzik, P.O., Finck, R., Bruggner, R.V., Melamed, R., Trejo, A., Ornatsky, O.I., et al. (2011). Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. Science *332*, 687–696.

Benner, C., Isoda, T., and Murre, C. (2015). New roles for DNA cytosine modification, eRNA, anchors, and superanchors in developing B cell progenitors. Proc. Natl. Acad. Sci. U. S. A. *112*, 12776–81.

Berland, R., and Wortis, H.H. (2002). Origins and Functions of B-1 Cells with Notes on the Role of CD5. Annu. Rev. Immunol. *20*, 253–300.

Bianchi, G., Anderson, K.C., Harris, N.L., and Sohani, A.R. (2014). The heavy chain diseases: clinical and pathologic features. Oncol. Williston Park N *28*, 45–53.

Bird, R.E., Hardman, K.D., Jacobson, J.W., Johnson, S., Kaufman, B.M., Lee, S.M., Lee, T., Pope, S.H., Riordan, G.S., and Whitlow, M. (1988). Single-chain antigen-binding proteins. Science *242*, 423–426.

Birshtein, B.K. (2012). The role of CTCF binding sites in the 3′ immunoglobulin heavy chain regulatory region. Front. Genet. *3*, 27.

Birshtein, B.K. (2014). Epigenetic Regulation of Individual Modules of the immunoglobulin heavy chain locus 3′ Regulatory Region. Front. Immunol. *5*, 163.

Blanco, N.D., Kruse, K., Erdmann, T., Staiger, A.M., Ott, G., Lenz, G., and Vaquerizas, J.M. (2018). Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. BioRxiv 372789.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. *2008*, P10008.

von Boehmer, H., and Melchers, F. (2010). Checkpoints in lymphocyte development and autoimmune disease. Nat. Immunol. *11*, 14–20.

Boekel, E. ten, Melchers, F., and Rolink, A.G. (1997). Changes in the VH Gene Repertoire of Developing Precursor B Lymphocytes in Mouse Bone Marrow Mediated by the Pre-B Cell Receptor. Immunity *7*, 357–368.

Boekel, E. ten, Melchers, F., and Rolink, A.G. (1998). Precursor B Cells Showing H Chain Allelic Inclusion Display Allelic Exclusion at the Level of Pre-B Cell Receptor Surface Expression. Immunity *8*, 199–207.

Boes, M., Esau, C., Fischer, Schmidt, T., Carroll, M., and Chen, J. (1998). Enhanced B-1 cell development, but impaired IgG antibody responses in mice deficient in secreted IgM. J. Immunol. Baltim. Md 1950 *160*, 4776–87.

Boes, M., Schmidt, T., Linkemann, K., Beaudette, B.C., Marshak-Rothstein, A., and Chen, J. (2000). Accelerated development of IgG autoantibodies and autoimmune disease in the absence of secreted IgM. Proc. Natl. Acad. Sci. U. S. A. *97*, 1184–9.

Bole, D.G., Hendershot, L.M., and Kearney, J.F. (1986). Posttranslational association of immunoglobulin heavy chain binding protein with nascent heavy chains in nonsecreting and secreting hybridomas. J. Cell Biol. *102*, 1558–1566.

Bolland, D.J., Wood, A.L., Johnston, C.M., Bunting, S.F., Morgan, G., Chakalova, L., Fraser, P.J., and Corcoran, A.E. (2004). Antisense intergenic transcription in V(D)J recombination. Nat. Immunol. *5*, 630–637.

Bolland, D.J., Wood, A.L., Afshar, R., Featherstone, K., Oltz, E.M., and Corcoran, A.E. (2007). Antisense Intergenic Transcription Precedes Igh D-to-J Recombination and Is Controlled by the Intronic Enhancer Eμ. Mol. Cell. Biol. *27*, 5523–5533.

Bolland, D.J., King, M.R., Reik, W., Corcoran, A.E., and Krueger, C. (2013). Robust 3D DNA FISH using directly labeled probes. J. Vis. Exp. JoVE.

Bolland, D.J., Koohy, H., Wood, A.L., Matheson, L.S., Krueger, F., Stubbington, M.J.T., Baizan-Edge, A., Chovanec, P., Stubbs, B.A., Tabbada, K., et al. (2016). Two Mutually Exclusive Local Chromatin States Drive Efficient V(D)J Recombination. Cell Rep. *15*, 2475–2487.

Bolotin, D.A., Mamedov, I.Z., Britanova, O.V., Zvyagin, I.V., Shagin, D., Ustyugova, S.V., Turchaninova, M.A., Lukyanov, S., Lebedev, Y.B., and Chudakov, D.M. (2012). Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. Eur. J. Immunol. *42*, 3073–83.

Bolotin, D.A., Shugay, M., Mamedov, I.Z., Putintseva, E.V., Turchaninova, M.A., Zvyagin, I.V., Britanova, O.V., and Chudakov, D.M. (2013). MiTCR: software for T-cell receptor sequencing data analysis. Nat. Methods *10*, 813–4.

Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., and Chudakov, D.M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods *12*, 380–1.

Bowen, A., and Corcoran, A. (2008). How chromatin remodelling allows shuffling of immunoglobulin heavy chain genes.

Brons, I.G.M., Smithers, L.E., Trotter, M.W.B., Rugg-Gunn, P., Sun, B., Lopes, S.M.C. de S., Howlett, S.K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R.A., et al. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature *448*, 191–195.

Brüggemann, M., Caskey, H., Teale, C., Waldmann, H., Williams, G., Surani, M., and Neuberger, M. (1989). A repertoire of monoclonal antibodies with human heavy chains from transgenic mice. Proc. Natl. Acad. Sci. U. S. A. *86*, 6709–13.

Brüggemann, M., Smith, J.A., Osborn, M.J., Corcos, D., Zou, X., Nguyen, V., and Muyldermans, S. (2006). Heavy-chain-only antibody expression and B-cell development in the mouse. Crit. Rev. Immunol. *26*, 377–90.

Brüggemann, M., Zou, X., Matheson, L., and Osborn, M. (2010). H-Chain-only antibodies.

Bryda, E.C., and Bauer, B.A. (2010). 20. A Restriction Enzyme-PCR-Based Technique to Determine Transgene Insertion Sites. Methods Mol. Biol. Clifton NJ *597*, 287–299.

Bulger, M., and Groudine, M. (2011). Functional and Mechanistic Diversity of Distal Transcription Enhancers. Cell *144*, 327–339.

Busslinger, M. (2004). Transcriptional control of early B cell development. Annu. Rev. Immunol. *22*, 55–79.

Cairns, J., Freire-Pritchett, P., Wingett, S.W., Várnai, C., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.-M.M., Osborne, C., et al. (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. *17*, 127.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661–667.

Carmack, C., Shinton, S., Hayakawa, K., and Hardy, R. (1990). Rearrangement and selection of VH11 in the Ly-1 B cell lineage. J. Exp. Med. *172*, 371–4.

Casola, S., Otipoby, K.L., Alimzhanov, M., Humme, S., Uyttersprot, N., Kutok, J.L., Carroll, M.C., and Rajewsky, K. (2004). B cell receptor signal strength determines B cell fate. Nat. Immunol. *5*, 317–27.

Chames, P., Van Regenmortel, M., Weiss, E., and Baty, D. (2009). Therapeutic antibodies: successes, limitations and hopes for the future. Br. J. Pharmacol. *157*, 220–233.

Chang, H.H.Y., Pannunzio, N.R., Adachi, N., and Lieber, M.R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat. Rev. Mol. Cell Biol. *18*, 495–506.

Chaudhary, N., and Wesemann, D.R. (2018). Analyzing Immunoglobulin Repertoires. Front. Immunol. *9*.

Chaumeil, J., Micsinai, M., Ntziachristos, P., Roth, D.B., Aifantis, I., Kluger, Y., Deriano, L., and Skok, J.A. (2013). The RAG2 C-terminus and ATM protect genome integrity by controlling antigen receptor gene cleavage. Nat. Commun. *4*, 2231.

Chen, W.V., and Maniatis, T. (2013). Clustered protocadherins. Dev. Camb. Engl. *140*, 3297–3302.

Choi, N.M., Loguercio, S., Verma-Gaur, J., Degner, S.C., Torkamani, A., Su, A.I., Oltz, E.M., Artyomov, M., and Feeney, A.J. (2013a). Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. J. Immunol. Baltim. Md 1950 *191*, 2393–402.

Choi, S.-C., Wang, H., Tian, L., Murakami, Y., Shin, D.-M., Borrego, F., Morse, H.C., and Coligan, J.E. (2013b). Mouse IgM Fc Receptor, FCMR, Promotes B Cell Development and Modulates Antigen-Driven Immune Responses. J. Immunol. *190*, 987–996.

Choi, Y.S., Dieter, J.A., Rothaeusler, K., Luo, Z., and Baumgarth, N. (2012). B-1 cells in the bone marrow are a significant source of natural IgM. Eur. J. Immunol. *42*, 120–129.

Chovanec, P., Bolland, D.J., Matheson, L.S., Wood, A.L., Krueger, F., Andrews, S., and Corcoran, A.E. (2018). Unbiased quantification of immunoglobulin diversity at the DNA level with VDJ-seq. Nat. Protoc. *13*, 1232–1252.

Chowdhury, D., and Sen, R. (2003). Transient IL-7/IL-7R signaling provides a mechanism for feedback inhibition of immunoglobulin heavy chain gene rearrangements. Immunity *18*, 229–241.

Christ, D., Famm, K., and Winter, G. (2007). Repertoires of aggregation-resistant human antibody domains. Protein Eng. Des. Sel. *20*, 413–416.

Collier, A.J., and Rugg-Gunn, P.J. (2018). Identifying Human Naïve Pluripotent Stem Cells − Evaluating State-Specific Reporter Lines and Cell-Surface Markers. BioEssays *40*, 1700239.

Collier, A.J., Panula, S.P., Schell, J.P., Chovanec, P., Reyes, A.P., Petropoulos, S., Corcoran, A.E., Walker, R., Douagi, I., Lanner, F., et al. (2017). Comprehensive Cell Surface Protein Profiling Identifies Specific Markers of Human Naive and Primed Pluripotent States. Cell Stem Cell *20*, 874–890.e7.

Conley, M., and Burrows, P.D. (2010). Plugging the leaky pre-B cell receptor. J. Immunol. Baltim. Md 1950 *184*.

Cooper, B.A., Sawai, C.M., Sicinska, E., Powers, S.E., Sicinski, P., Clark, M.R., and Aifantis, I. (2006). A unique function for cyclin D3 in early B cell development. Nat. Immunol. *7*.

Corcoran, A.E. (2010). The epigenetic role of non-coding RNA transcription and nuclear organization in immunoglobulin repertoire generation. Semin. Immunol. *22*, 353–361.

Corcos, D. (1990). Oncogenic potential of the B-cell antigen receptor and its relevance to heavy chain diseases and other B-cell neoplasias: A new model. Res. Immunol. *141*, 543–553.

Corcos, D. (2010). Immunoglobulin transport in the absence of light chains. Trends Biochem. Sci. *35*, 593.

Corcos, D., Iglesias, A., Dunda, O., and Jami, J. (1991). Allelic exclusion in transgenic mice expressing a heavy chain disease-like human μ protein. Eur. J. Immunol. *21*, 2711–2716.

Corcos, D., Dunda, O., Butor, C., Cesbron, J.-Y., Lorès, P., Bucchini, D., and Jami, J. (1995). Pre-B-cell development in the absence of λ5 in transgenic mice expressing a heavy-chain disease protein. Curr. Biol. *5*, 1140–1148.

Corcos, D., Grandien, A., Vazquez, A., Dunda, O., Lorès, P., and Bucchini, D. (2001). Expression of a V Region-Less B Cell Receptor Confers a Tolerance-Like Phenotype on Transgenic B Cells. J. Immunol. *166*, 3083–3089.

Corcos, D., Osborn, M.J., Matheson, L.S., Santos, F., Zou, X., Smith, J.A., Morgan, G., Hutchings, A., Hamon, M., Oxley, D., et al. (2010). Immunoglobulin aggregation leading to Russell body formation is prevented by the antibody light chain. Blood *115*.

Corcos, D., Osborn, M.J., and Matheson, L.S. (2011). B-cell receptors and heavy chain diseases: guilty by association? Blood *117*, 6991–6998.

Cowell, L.G., Davila, M., Kepler, T.B., and Kelsoe, G. (2002). Identification and utilization of arbitrary correlations in models of recombination signal sequences. Genome Biol. *3*, research0072.1.

Cowell, L.G., Davila, M., Yang, K., Kepler, T.B., and Kelsoe, G. (2003). Prospective Estimation of Recombination Signal Efficiency and Identification of Functional Cryptic Signals in the Genome by Statistical Modeling. J. Exp. Med. *197*, 207–220.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. *2*, 292–301.

Cremer, T., Cremer, M., Dietzel, S., Müller, S., Solovei, I., and Fakan, S. (2006). Chromosome territories – a functional nuclear landscape. Curr. Opin. Cell Biol. *18*, 307–316.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: A Sequence Logo Generator. Genome Res. *14*, 1188–1190.

Cruz-Molina, S., Respuela, P., Tebartz, C., Kolovos, P., Nikolic, M., Fueyo, R., van Ijcken, W.F.J., Grosveld, F., Frommolt, P., Bazzi, H., et al. (2017). PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. Cell Stem Cell *20*, 689–705.e9.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Syst. *1695*, 1–9.

Das, S., Nozawa, M., Klein, J., and Nei, M. (2008). Evolutionary dynamics of the immunoglobulin heavy chain variable region genes in vertebrates. Immunogenetics *60*, 47–55.

Davies, J., and Riechmann, L. (1994). 'Camelising' human antibody fragments: NMR studies on VH domains. FEBS Lett. *339*, 285–290.

Decker, D.J., Boyle, N.E., Koziol, J.A., and Klinman, N.R. (1991). The expression of the Ig H chain repertoire in developing bone marrow B lineage cells. J. Immunol. *146*, 350–361.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. Science *295*, 1306–1311.

DeKosky, B.J., Ippolito, G.C., Deschner, R.P., Lavinder, J.J., Wine, Y., Rawlings, B.M., Varadarajan, N., Giesecke, C., Dörner, T., Andrews, S.F., et al. (2013). High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat. Biotechnol. *31*, 166–169.

DeKosky, B.J., Lungu, O.I., Park, D., Johnson, E.L., Charab, W., Chrysostomou, C., Kuroda, D., Ellington, A.D., Ippolito, G.C., Gray, J.J., et al. (2016). Large-scale sequence and structural

comparisons of human naive and antigen-experienced antibody repertoires. Proc. Natl. Acad. Sci. U. S. A. *113*, E2636–45.

Deng, C., Daley, T., and Smith, A. (2015). Applications of species accumulation curves in large-scale biological data analysis. Quant. Biol. *3*, 135–144.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., and Blobel, G.A. (2012). Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. Cell *149*, 1233–1244.

Desiderio, S.V., Yancopoulos, G.D., Paskind, M., Thomas, E., Boss, M.A., Landau, N., Alt, F.W., and Baltimore, D. (1984). Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells. Nature *311*, 752–755.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Dudgeon, K., Famm, K., and Christ, D. (2009). Sequence determinants of protein aggregation in human VH domains. Protein Eng. Des. Sel.

Dunn-Walters, D., Townsend, C., Sinclair, E., and Stewart, A. (2018). Immunoglobulin gene analysis as a tool for investigating human immune responses. Immunol. Rev. *284*, 132–147.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Eberle, A.B., Herrmann, K., Jäck, H.-M.M., and Mühlemann, O. (2009). Equal transcription rates of productively and nonproductively rearranged immunoglobulin mu heavy chain alleles in a pro-B cell line. RNA N. Y. N *15*, 1021–8.

Ecker, D.M., Jones, S.D., and Levine, H.L. (2015). The therapeutic monoclonal antibody market. MAbs *7*, 9–14.

Ehlich, A., Martin, V., Müller, W., and Rajewsky, K. (1994). Analysis of the B-cell progenitor compartment at the level of single cells. Curr. Biol. *4*, 573–583.

Ehrenstein, M.R., and Notley, C.A. (2010). The importance of natural IgM: scavenger, protector and regulator. Nat. Rev. Immunol. *10*, 778–86.

Ehrenstein, M., O'Keefe, T., Davies, S., and Neuberger, M. (1998). Targeted gene disruption reveals a role for natural secretory IgM in the maturation of the primary immune response. Proc. Natl. Acad. Sci. U. S. A. *95*, 10089–93.

Ehrenstein, M.R., Cook, H.T., and Neuberger, M.S. (2000). Deficiency in Serum Immunoglobulin (Ig)m Predisposes to Development of Igg Autoantibodies. J. Exp. Med. *191*, 1253–1258.

Eijsbouts, C., Burren, O., Newcombe, P., and Wallace, C. (2018). Fine mapping chromatin contacts in capture Hi-C data. BioRxiv 243642.

Eils, R., Dietzel, S., Bertin, E., Schröck, E., Speicher, M.R., Ried, T., Robert-Nicoud, M., Cremer, C., and Cremer, T. (1996). Three-dimensional reconstruction of painted human interphase chromosomes: active and inactive X chromosome territories have similar volumes but differ in shape and surface structure. J. Cell Biol. *135*, 1427–1440.

Ellgaard, L., and Helenius, A. (2003). Quality control in the endoplasmic reticulum. Nat. Rev. Mol. Cell Biol. *4*, 181–91.

Ellis B, Haaland P, Hahne F, Le Meur N, Gopalakrishnan N, Spidlen J, and Jiang M (2017). flowCore: Basic structures for flow cytometry data.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. *12*, 2478–2492.

Ewert, S., Cambillau, C., Conrath, K., and Plückthun, A. (2002). Biophysical Properties of Camelid VHH Domains Compared to Those of Human VH3 Domains†. Biochemistry (Mosc.) *41*, 3628–3636.

Featherstone, K., Wood, A.L., Bowen, A.J., and Corcoran, A.E. (2010). The Mouse Immunoglobulin Heavy Chain V-D Intergenic Sequence Contains Insulators That May Regulate Ordered V(D)J Recombination. J. Biol. Chem. *285*, 9327–9338.

Feeney, A.J., Goebel, P., and Espinoza, C.R. (2004). Many levels of control of V gene rearrangement frequency. Immunol. Rev. *200*, 44–56.

Feige, M.J., Groscurth, S., Marcinowski, M., Shimizu, Y., Kessler, H., Hendershot, L.M., and Buchner, J. (2009). An unfolded CH1 domain controls the assembly and secretion of IgG antibodies. Mol. Cell *34*, 569–79.

Ferry, H., Potter, P.K., Crockford, T.L., Nijnik, A., Ehrenstein, M.R., Walport, M.J., Botto, M., and Cornall, R.J. (2007). Increased Positive Selection of B1 Cells and Reduced B Cell Tolerance to Intracellular Antigens in c1q-Deficient Mice. J. Immunol. *178*, 2916–2922.

Fishwild, D., O'Donnell, S., Bengoechea, T., Hudson, D., Harding, F., Bernhard, S., Jones, D., Kay, R., Higgins, K., Schramm, S., et al. (1996). High-avidity human IgG kappa monoclonal antibodies from a novel strain of minilocus transgenic mice. Nat. Biotechnol. *14*, 845–51.

Flajnik, M.F. (2002). Comparative analyses of immunoglobulin genes: surprises and portents. Nat. Rev. Immunol. *2*, 688–98.

Flajnik, M.F., and Kasahara, M. (2010). Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat. Rev. Genet. *11*, 47–59.

Freire-Pritchett, P., Schoenfelder, S., Várnai, C., Wingett, S.W., Cairns, J., Collier, A.J., García-Vílchez, R., Furlan-Magaril, M., Osborne, C.S., Fraser, P., et al. (2017). Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. ELife *6*, e21926.

Fu, G.K., Hu, J., Wang, P.-H., and Fodor, S.P. (2011). Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc. Natl. Acad. Sci. *108*, 9026–9031.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. *15*, 2038–2049.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Gaëta, B.A., Malming, H.R., Jackson, K.J.L., Bain, M.E., Wilson, P., and Collins, A.M. (2007). iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. Bioinformatics *23*, 1580–1587.

Galson, J.D., Pollard, A.J., Trück, J., and Kelly, D.F. (2014). Studying the antibody repertoire after vaccination: practical applications. Trends Immunol. *35*, 319–331.

Garrett, F.E., Emelyanov, A.V., Sepulveda, M.A., Flanagan, P., Volpi, S., Li, F., Loukinov, D., Eckhardt, L.A., Lobanenkov, V.V., and Birshtein, B.K. (2005). Chromatin Architecture near a Potential 3′ End of the Igh Locus Involves Modular Regulation of Histone Modifications during B-Cell Development and In Vivo Occupancy at CTCF Sites. Mol. Cell. Biol. *25*, 1511–1525.

Gassen, S., Callebaut, B., Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A *87*, 636–645.

Geisberger, R., Crameri, R., and Achatz, G. (2003). Models of signal transduction through the B-cell antigen receptor. Immunology *110*, 401–410.

Gel, B., Serra, E., and Hancock, J. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics *33*, 3088–3090.

Genst, E.D., Saerens, D., Muyldermans, S., and Conrath, K. (2006). Antibody repertoire development in camelids. Dev. Comp. Immunol. *30*, 187–98.

Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H., and Quake, S.R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat. Biotechnol. *32*, 158–68.

Geraldes, P., Rebrovich, M., Herrmann, K., Wong, J., Jäck, H.-M.M., Wabl, M., and Cascalho, M. (2007). Ig heavy chain promotes mature B cell survival in the absence of light chain. J. Immunol. Baltim. Md 1950 *179*, 1659–68.

Goloborodko, A., Marko, J.F., and Mirny, L.A. (2016). Chromosome Compaction by Active Loop Extrusion. Biophys. J. *110*, 2162–2168.

Gonen, N., Futtner, C.R., Wood, S., Garcia-Moreno, S.A., Salamone, I.M., Samson, S.C., Sekido, R., Poulat, F., Maatouk, D.M., and Lovell-Badge, R. (2018). Sex reversal following deletion of a single distal enhancer of Sox9. Science eaas9408.

Green, L.L., Hardy, M.C., Maynard-Currie, C.E., Tsuda, H., Louie, D.M., Mendez, M.J., Abderrahim, H., Noguchi, M., Smith, D.H., Zeng, Y., et al. (1994). Antigen–specific human monoclonal antibodies from mice engineered with human Ig heavy and light chain YACs. Nat. Genet. *7*, 13–21.

Greiff, V., Miho, E., Menzel, U., and Reddy, S.T. (2015). Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. Trends Immunol. *36*, 738–749.

Greiff, V., Menzel, U., Miho, E., Weber, C., Riedel, R., Cook, S., Valai, A., Lopes, T., Radbruch, A., Winkler, T.H., et al. (2017a). Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. Cell Rep. *19*, 1467–1478.

Greiff, V., Weber, C.R., Palme, J., Bodenhofer, U., Miho, E., Menzel, U., and Reddy, S.T. (2017b). Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. J. Immunol. *199*, 2985–2997.

Grimsholm, O., Ren, W., Bernardi, A.I., Chen, H., Park, G., Camponeschi, A., Chen, D., Bergmann, B., Höök, N., Andersson, S., et al. (2015). Absence of surrogate light chain results in spontaneous autoreactive germinal centres expanding VH81X-expressing B cells. Nat. Commun. *6*, 7077.

Gu, W., Crawford, E.D., O'Donovan, B.D., Wilson, M.R., Chow, E.D., Retallack, H., and DeRisi, J.L. (2016). Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol. *17*, 41.

Guhr, A., Kobold, S., Seltmann, S., Wulczyn, A.E.M.S., Kurtz, A., and Löser, P. (2018). Recent Trends in Research with Human Pluripotent Stem Cells: Impact of Research and Use of Cell Lines in Experimental Research and Clinical Trials. Stem Cell Rep. *11*, 485–496.

Guloglu, B.F., and Roman, C.A. (2006). Precursor B Cell Receptor Signaling Activity Can Be Uncoupled from Surface Expression. J. Immunol. *176*, 6862–6872.

Guo, C., Gerasimova, T., Hao, H., Ivanova, I., and Chakraborty, T. (2011a). Two forms of loops generate the chromatin conformation of the immunoglobulin heavy-chain gene locus.

Guo, C., Yoon, H.S., Franklin, A., Jain, S., Ebert, A., Cheng, H.-L., Hansen, E., Despo, O., Bossen, C., Vettermann, C., et al. (2011b). CTCF-binding elements mediate control of V(D)J recombination. Nature *477*, 424–430.

Guo, G., von Meyenn, F., Santos, F., Chen, Y., Reik, W., Bertone, P., Smith, A., and Nichols, J. (2016). Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass. Stem Cell Rep. *6*, 437–446.

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J., et al. (2014). The DNA methylation landscape of human early embryos. Nature *511*, 606–610.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–10.

Gupta, N.T., Heiden, V., A, J., Uduman, M., Gadala-Maria, D., Yaari, G., and Kleinstein, S.H. (2015). Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics *31*, 3356–3358.

Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell *169*, 693–707.e14.

Haas, I.G., and Wabl, M. (1983). Immunoglobulin heavy chain binding protein. Nature *306*, 387–389.

Hahne, F., and Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. In Statistical Genomics: Methods and Protocols, E. Mathé, and S. Davis, eds. (New York, NY: Springer New York), pp. 335–351.

Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hammers, C., Songa, E.B., Bendahman, N., and Hammers, R. (1993). Naturally occurring antibodies devoid of light chains. Nature *363*, 446–448.

Han, C., Choi, E., Park, I., Lee, B., Jin, S., Kim, D.H. o, Nishimura, H., and Cho, C. (2009). Comprehensive analysis of reproductive ADAMs: relationship of ADAM4 and ADAM6 with an ADAM complex required for fertilization in mice. Biol. Reprod. *80*, 1001–8.

Hardy, R., Carmack, C., Shinton, S., Kemp, J., and Hayakawa, K. (1991). Resolution and characterization of pro-B and pre-pro-B cell stages in normal mouse bone marrow. J. Exp. Med. *173*.

Hardy, R.R., Kincade, P.W., and Dorshkind, K. (2007). The protean nature of cells in the B lymphocyte lineage. Immunity *26*, 703–14.

Harwood, N.E., and Batista, F.D. (2010). Early events in B cell activation. Annu. Rev. Immunol. *28*, 185–210.

Hasegawa, H., Woods, C.E., Kinderman, F., He, F., and Lim, A.C. (2014). Russell body phenotype is preferentially induced by IgG mAb clones with high intrinsic condensation propensity: Relations between the biosynthetic events in the ER and solution behaviors in vitro. MAbs *6*, 1518–1532.

Hayakawa, K., Hardy, R.R., Herzenberg, L.A., and Herzenberg, L.A. (1985). Progenitors for Ly-1 B cells are distinct from progenitors for other B cells. J. Exp. Med. *161*, 1554–68.

Hayakawa, K., Hardy, R.R., Stall, A.M., Herzenberg, L.A., and Herzenberg, L.A. (1986). Immunoglobulin-bearing B cells reconstitute and maintain the murine Ly-1 B cell lineage. Eur. J. Immunol. *16*, 1313–1316.

Hayashi, K., Nittono, R., Okamoto, N., Tsuji, S., Hara, Y., Goitsuka, R., and Kitamura, D. (2000). The B cell-restricted adaptor BASH is required for normal development and antigen receptor-mediated activation of B cells. Proc. Natl. Acad. Sci. U. S. A. *97*, 2755–2760.

Hayday, A.C., Gillies, S.D., Saito, H., Wood, C., Wiman, K., Hayward, W.S., and Tonegawa, S. (1984). Activation of a translocated human c-myc gene by an enhancer in the immunoglobulin heavy-chain locus. Nature *307*, 334–340.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Heltemes, L.M., and Manser, T. (2002). Level of B Cell Antigen Receptor Surface Expression Influences Both Positive and Negative Selection of B Cells During Primary Development. J. Immunol. *169*, 1283–1292.

Hendershot, L., Bole, D., Köhler, G., and Kearney, J.F. (1987). Assembly and secretion of heavy chains that do not associate posttranslationally with immunoglobulin heavy chain-binding protein. J. Cell Biol. *104*, 761–7.

Herzog, S., Reth, M., and Jumaa, H. (2009). Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling. Nat. Rev. Immunol. *9*, 195–205.

Hess, J., Werner, A., Wirth, T., Melchers, F., Jäck, H.-M., and Winkler, T.H. (2001). Induction of pre-B cell proliferation after de novo synthesis of the pre-B cell receptor. Proc. Natl. Acad. Sci. *98*, 1745–1750.

Hewitt, S.L., Farmer, D., Marszalek, K., Cadera, E., Liang, H.-E.E., Xu, Y., Schlissel, M.S., and Skok, J.A. (2008). Association between the Igk and Igh immunoglobulin loci mediated by the 3' Igk enhancer induces "decontraction" of the Igh locus in pre-B cells. Nat. Immunol. *9*, 396–404.

Hewitt, S.L., Yin, B., Ji, Y., Chaumeil, J., Marszalek, K., Tenthorey, J., Salvagiotto, G., Steinel, N., Ramsey, L.B., Ghysdael, J., et al. (2009). RAG-1 and ATM coordinate monoallelic recombination and nuclear positioning of immunoglobulin loci. Nat. Immunol.

Holliger, P., Prospero, T., and Winter, G. (1993). "Diabodies": small bivalent and bispecific antibody fragments. Proc. Natl. Acad. Sci. *90*, 6444–6448.

Holwerda, S.J.B., and de Laat, W. (2013). CTCF: the protein, the binding partners, the binding sites and their chromatin loops. Philos. Trans. R. Soc. B Biol. Sci. *368*.

Holwerda, S.J., van de Werken, H.J., Ribeiro de Almeida, C., Bergen, I.M., de Bruijn, M.J., Verstegen, M.J., Simonis, M., Splinter, E., Wijchers, P.J., Hendriks, R.W., et al. (2013). Allelic exclusion of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells. Nucleic Acids Res. *41*, 6905–6916.

Hou, C., Li, L., Qin, Z.S., and Corces, V.G. (2012). Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. Mol. Cell *48*, 471–484.

Hou, D., Ying, T., Wang, L., Chen, C., Lu, S., Wang, Q., Seeley, E., Xu, J., Xi, X., Li, T., et al. (2016a). Immune Repertoire Diversity Correlated with Mortality in Avian Influenza A (H7N9) Virus Infected Patients. Sci. Rep. *6*, 33843.

Hou, D., Chen, C., Seely, E.J., Chen, S., and Song, Y. (2016b). High-Throughput Sequencing-Based Immune Repertoire Study during Infectious Disease. Front. Immunol. *7*, 336.

Hozák, P., Hassan, A.B., Jackson, D.A., and Cook, P.R. (1993). Visualization of replication factories attached to a nucleoskeleton. Cell *73*, 361–373.

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

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods *12*, 115–121.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95.

Igarashi, H., Gregory, S.C., Yokota, T., Sakaguchi, N., and Kincade, P.W. (2002). Transcription from the RAG1 locus marks the earliest lymphocyte progenitors in bone marrow. Immunity *17*, 117–30.

Imakaev, M., Fudenberg, G., McCord, R., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods *9*, 999–1003.

Islam, S., Zeisel, A., Joost, S., Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.

Ivanov, I.I., Schelonka, R.L., Zhuang, Y., Gartland, G.L., Zemlin, M., and Schroeder, H.W. (2005). Development of the Expressed Ig CDR-H3 Repertoire Is Marked by Focusing of Constraints in Length, Amino Acid Use, and Charge That Are First Established in Early B Cell Progenitors. J. Immunol. *174*, 7773–7780.

Jackson, K.J., Kidd, M.J., Wang, Y., and Collins, A.M. (2013). The Shape of the Lymphocyte Receptor Repertoire: Lessons from the B Cell Receptor. Front. Immunol. *4*, 263.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLOS ONE *9*, e98679.

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369–1384.e19.

Jayaram, N., Bhowmick, P., and Martin, A.C.R. (2012). Germline VH/VL pairing in antibodies. Protein Eng. Des. Sel. *25*, 523–530.

Jespers, L., Schon, O., Famm, K., and Winter, G. (2004). Aggregation-resistant domain antibodies selected on phage by heat denaturation. Nat. Biotechnol. *22*, 1161–1165.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I., et al. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell *18*, 262–275.

Johnson, K., Hashimshony, T., Sawai, C.M., Pongubala, J.M.R., Skok, J.A., Aifantis, I., and Singh, H. (2008). Regulation of Immunoglobulin Light-Chain Recombination by the Transcription Factor IRF-4 and the Attenuation of Interleukin-7 Signaling. Immunity *28*, 335–345.

Johnston, C.M., Wood, A.L., Bolland, D.J., and Corcoran, A.E. (2006). Complete Sequence Assembly and Characterization of the C57BL/6 Mouse Ig Heavy Chain V Region. J. Immunol. *176*, 4221–4234.

Joshi, O., Wang, S.-Y., Kuznetsova, T., Atlasi, Y., Peng, T., Fabre, P.J., Habibi, E., Shaik, J., Saeed, S., Handoko, L., et al. (2015). Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency. Cell Stem Cell *17*, 748–757.

Jumaa, H., Wollscheid, B., Mitterer, M., Wienands, J., Reth, M., and Nielsen, P.J. (1999). Abnormal development and function of B lymphocytes in mice deficient for the signaling adaptor protein SLP-65. Immunity *11*, 547–554.

Kaloff, C.R., and Haas, I.G. (1995). Coordination of immunoglobulin chain folding and immunoglobulin chain assembly is essential for the formation of functional IgG. Immunity *2*, 629–637.

Kaplinsky, J., Li, A., Sun, A., Coffre, M., Koralov, S.B., and Arnaout, R. (2014). Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. Proc. Natl. Acad. Sci. *111*, E2622–E2629.

Kawano, Y., Ouchida, R., Wang, J.-Y., Yoshikawa, S., Yamamoto, M., Kitamura, D., and Karasuyama, H. (2012). A Novel Mechanism for the Autonomous Termination of Pre-B Cell Receptor Expression via Induction of Lysosome-Associated Protein Transmembrane 5. Mol. Cell. Biol. *32*, 4462–4471.

Keenan, R.A., Riva, A.D., Corleis, B., Hepburn, L., Licence, S., Winkler, T.H., and Mårtensson, I.-L.L. (2008). Censoring of autoreactive B cell development by the pre-B cell receptor. Science *321*, 696–9.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kind, J., Pagie, L., Vries, S.S. de, Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., Graaf, C.A. de, Amendola, M., et al. (2015). Genome-wide maps of nuclear lamina interactions in single human cells. Cell *163*, 134–47.

Kirkham, P.M., Mortari, F., Newton, J.A., and Schroeder, H.W. (1992). Immunoglobulin VH clan and family identity predicts variable domain structure and may influence antigen binding. EMBO J. *11*, 603–609.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods *9*, 72–74.

Knight, A.M., Lucocq, J.M., Prescott, A.R., Ponnambalam, S., and Watts, C. (1997). Antigen endocytosis and presentation mediated by human membrane IgG1 in the absence of the Igα/Igβ dimer. EMBO J. *16*, 3842–3850.

Kodaira, M., Kinashi, T., Umemura, I., Matsuda, F., Noma, T., Ono, Y., and Honjo, T. (1986). Organization and evolution of variable region genes of the human immunoglobulin heavy chain. J. Mol. Biol. *190*, 529–541.

Köhler, G., and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. Nature *256*, 495–497.

Köhler, F., Hug, E., Eschbach, C., Meixlsperger, S., Hobeika, E., Kofer, J., Wardemann, H., and Jumaa, H. (2008). Autoreactive B cell receptors mimic autonomous pre-B cell receptor signaling and induce proliferation of early B cells. Immunity *29*, 912–21.

Koohy, H., Bolland, D.J., Matheson, L.S., Schoenfelder, S., Stellato, C., Dimond, A., Várnai, C., Chovanec, P., Chessa, T., Denizot, J., et al. (2018). Genome organization and chromatin analysis identify transcriptional downregulation of insulin-like growth factor signaling as a hallmark of aging in developing B cells. Genome Biol. *19*, 126.

Kosak, S.T., Skok, J.A., Medina, K.L., Riblet, R., Beau, M.M.L., Fisher, A.G., and Singh, H. (2002). Subnuclear Compartmentalization of Immunoglobulin Loci During Lymphocyte Development. Science *296*, 158–162.

Kottmann, A.H., Zevnik, B., Welte, M., Nielsen, P.J., and Köhler, G. (1994). A second promoter and enhancer element within the immunoglobulin heavy chain locus. Eur. J. Immunol. *24*, 817–821.

Kruskal, J.B., and Wish, M. (1978). Multidimensional Scaling (SAGE).

Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res.

Kubagawa, H., Skopnik, C.M., Zimmermann, J., Durek, P., Chang, H.-D., Yoo, E., Bertoli, L.F., Honjo, K., and Radbruch, A. (2017). Authentic IgM Fc Receptor (FcμR). In IgM and Its Receptors and Binding Proteins, H. Kubagawa, and P.D. Burrows, eds. (Cham: Springer International Publishing), pp. 25–45.

Laat, W. de, and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. Nature *502*, 499–506.

Lam, K.P., and Rajewsky, K. (1999). B cell antigen receptor specificity and surface density together determine B-1 versus B-2 cell development. J. Exp. Med. *190*, 471–477.

Lam, K.P., Kühn, R., and Rajewsky, K. (1997). In vivo ablation of surface immunoglobulin on mature B cells by inducible gene targeting results in rapid cell death. Cell *90*, 1073–83.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Larimore, K., McCormick, M.W., Robins, H.S., and Greenberg, P.D. (2012). Shaping of human germline IgH repertoires revealed by deep sequencing. J. Immunol. *189*, 3221–3230.

Lassmann, T., Frings, O., and Sonnhammer, E.L.L. (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res. *37*, 858–865.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLOS Comput. Biol. *9*, e1003118.

Lee, J., and Desiderio, S. (1999). Cyclin A/CDK2 regulates V(D)J recombination by coordinating RAG-2 accumulation and DNA repair. Immunity *11*, 771–781.

Lee, E.-C., Liang, Q., Ali, H., Bayliss, L., Beasley, A., Bloomfield-Gerdes, T., Bonoli, L., Brown, R., Campbell, J., Carpenter, A., et al. (2014). Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. Nat. Biotechnol. *32*, 356–363.

Lee, J.-W., Chen, Z., Geng, H., Xiao, G., Park, E., Parekh, S., Kornblau, S.M., Melnick, A., Abbas, A., Paietta, E., et al. (2015). CD25 (IL2RA) Orchestrates Negative Feedback Control and Stabilizes Oncogenic Signaling Strength in Acute Lymphoblastic Leukemia. Blood *126*, 1434–1434.

Lee, Y.-K., Brewer, J.W., Hellman, R., and Hendershot, L.M. (1999). BiP and Immunoglobulin Light Chain Cooperate to Control the Folding of Heavy Chain and Ensure the Fidelity of Immunoglobulin Assembly. Mol. Biol. Cell *10*, 2209–2219.

Lefranc, M.P. (1998). IMGT (ImMunoGeneTics) locus on focus. A new section of Experimental and Clinical Immunogenetics. Exp. Clin. Immunogenet. *15*, 1–7.

Lefranc, M.-P. (2000). Nomenclature of the Human Immunoglobulin Genes. Curr. Protoc. Immunol. *40*, A.1P.1-A.1P.37.

Lefranc, M.-P. (2014). Immunoglobulins: 25 Years of Immunoinformatics and IMGT-ONTOLOGY. Biomolecules *4*, 1102–1139.

Lefranc, M.-P., and Lefranc, G. (2001). The Immunoglobulin FactsBook (Academic Press).

Leitzgen, K., Knittler, M.R., and Haas, I.G. (1997). Assembly of immunoglobulin light chains as a prerequisite for secretion. A model for oligomerization-dependent subunit folding. J. Biol. Chem. *272*, 3117–23.

Lennon, G.G., and Perry, R.P. (1985). Cμ-containing transcripts initiate heterogeneously within the IgH enhancer region and contain a novel 5′-nontranslatable exon. Nature *318*, 475–478.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

Li, B., Carey, M., and Workman, J.L. (2007). The Role of Chromatin during Transcription. Cell *128*, 707–719.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. Cell *148*, 84–98.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science *326*, 289–293.

Lin, S.G., Ba, Z., Du, Z., Zhang, Y., Hu, J., and Alt, F.W. (2016). Highly sensitive and unbiased approach for elucidating antibody repertoires. Proc. Natl. Acad. Sci. *113*, 7846–7851.

Little, A.J., Matthews, A., Oettinger, M., Roth, D.B., and Schatz, D.G. (2015). Chapter 2 - The Mechanism of V(D)J Recombination. In Molecular Biology of B Cells (Second Edition), F.W. Alt, T. Honjo, A. Radbruch, and M. Reth, eds. (London: Academic Press), pp. 13–34.

Loder, B., Mutschler, B., Ray, R.J., Paige, C.J., Sideras, P., Torres, R., Lamers, M.C., and Carsetti, R. (1999). B Cell Development in the Spleen Takes Place in Discrete Steps and Is Determined by the Quality of B Cell Receptor–Derived Signals. J. Exp. Med. *190*.

Lonberg, N., Taylor, L., Harding, F., Trounstine, M., Higgins, K., Schramm, S., Kuo, C., Mashayekh, R., Wymore, K., and McCabe, J. (1994). Antigen-specific human antibodies from mice comprising four distinct genetic modifications. Nature *368*, 856–9.

Longo, N.S., Rogosch, T., Zemlin, M., Zouali, M., and Lipsky, P.E. (2017). Mechanisms That Shape Human Antibody Repertoire Development in Mice Transgenic for Human Ig H and L Chain Loci. J. Immunol. Baltim. Md 1950.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Love, V.A., Lugo, G., Merz, D., and Feeney, A.J. (2000). Individual VH promoters vary in strength, but the frequency of rearrangement of those VH genes does not correlate with promoter strength nor enhancer-independence. Mol. Immunol. *37*, 29–39.

Lutz, J., Heideman, M.R., Roth, E., Berk, P. van den, Müller, W., Raman, C., Wabl, M., Jacobs, H., and Jäck, H.-M.M. (2011). Pro-B cells sense productive immunoglobulin heavy chain rearrangement irrespective of polypeptide production. Proc. Natl. Acad. Sci. U. S. A. *108*, 10644–9.

Ma, B., Osborn, M.J., Avis, S., Ouisse, L.-H.H., Ménoret, S., Anegon, I., Buelow, R., and Brüggemann, M. (2013). Human antibody expression in transgenic rats: comparison of chimeric IgH loci with human VH, D and JH but bearing different rat C-gene regions. J. Immunol. Methods *400–401*, 78–86.

Ma, S., Pathak, S., Trinh, L., and Lu, R. (2008). Interferon regulatory factors 4 and 8 induce the expression of Ikaros and Aiolos to down-regulate pre–B-cell receptor and promote cell-cycle withdrawal in pre–B-cell development. Blood *111*, 1396–1403.

Ma, Y., Pannicke, U., Schwarz, K., and Lieber, M.R. (2002). 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.

Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Macdonald, L.E., Karow, M., Stevens, S., Auerbach, W., Poueymirou, W.T., Yasenchak, J., Frendewey, D., Valenzuela, D.M., Giallourakis, C.C., Alt, F.W., et al. (2014). Precise and in situ genetic humanization of 6 Mb of mouse immunoglobulin genes. Proc. Natl. Acad. Sci. *111*, 5147–5152.

Mainville, C.A., Sheehan, K.M., Klaman, L.D., Giorgetti, C.A., Press, J.L., and Brodeur, P.H. (1996). Deletional mapping of fifteen mouse VH gene families reveals a common organization for three Igh haplotypes. J. Immunol. *156*, 1038–1046.

Malu, S., Malshetty, V., Francis, D., and Cortes, P. (2012). Role of non-homologous end joining in V(D)J recombination. Immunol. Res. *54*, 233–246.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Francis Stewart, A., Smith, A., et al. (2012). The Transcriptional and Epigenomic Foundations of Ground State Pluripotency. Cell *149*, 590–604.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10–12.

Martin, F., and Kearney, J.F. (2002). Marginal-zone B cells. Nat. Rev. Immunol. *2*, 323–35.

Martin, S.W., and Goodnow, C.C. (2002). Burst-enhancing role of the IgG membrane tail as a molecular determinant of memory. Nat. Immunol. *3*, 182–188.

Martin, V., Wu, Y.-C., Kipling, D., and Dunn-Walters, D. (2015). Ageing of the B-cell repertoire. Phil Trans R Soc B *370*, 20140237.

Martin, V.G., Wu, Y.-C.B., Townsend, C.L., Lu, G.H.C., O'Hare, J.S., Mozeika, A., Coolen, A.C.C., Kipling, D., Fraternali, F., and Dunn-Walters, D.K. (2016). Transitional B Cells in Early Human B Cell Development – Time to Revisit the Paradigm? Front. Immunol. *7*.

Matheson, L., and Elderkin, S. (2018). 13 - Polycomb Bodies. In Nuclear Architecture and Dynamics, C. Lavelle, and J.-M. Victor, eds. (Boston: Academic Press), pp. 297–320.

Matheson, L.S., and Corcoran, A.E. (2012). Local and global epigenetic regulation of V(D)J recombination. Curr. Top. Microbiol. Immunol. *356*, 65–89.

Matheson, L.S., Bolland, D.J., Chovanec, P., Krueger, F., Andrews, S., Koohy, H., and Corcoran, A. (2017). Local chromatin features including PU.1 and IKAROS binding and H3K4 methylation shape the repertoire of immunoglobulin kappa genes chosen for V(D)J recombination. Front. Immunol. *8*, 1550.

Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K., Hayashida, H., Miyata, T., and Honjo, T. (1998). The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J. Exp. Med. *188*, 2151–2162.

Mattioli, L., Anelli, T., Fagioli, C., Tacchetti, C., Sitia, R., and Valetti, C. (2006). ER storage diseases: a role for ERGIC-53 in controlling the formation and shape of Russell bodies. J. Cell Sci. *119*, 2532–2541.

Maus, M.V., and June, C.H. (2016). Making Better Chimeric Antigen Receptors for Adoptive T-cell Therapy. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *22*, 1875–84.

McCafferty, J., Griffiths, A.D., Winter, G., and Chiswell, D.J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. Nature *348*, 552–554.

McDaniel, J.R., DeKosky, B.J., Tanno, H., Ellington, A.D., and Georgiou, G. (2016). Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. Nat. Protoc. *11*, 429–42.

Medvedovic, J., Ebert, A., Tagoh, H., Tamir, I.M., Schwickert, T.A., Novatchkova, M., Sun, Q., Veld, P.J., Guo, C., Yoon, H.S., et al. (2013). Flexible long-range loops in the VH gene region of the Igh locus facilitate the generation of a diverse antibody repertoire. Immunity *39*, 229–44.

Meixlsperger, S., Köhler, F., Wossning, T., Reppel, M., Müschen, M., and Jumaa, H. (2007). Conventional light chains inhibit the autonomous signaling capacity of the B cell receptor. Immunity *26*, 323–33.

Melchers, F. (1999). Fit for life in the immune system? Surrogate L chain tests H chains that test L chains. Proc. Natl. Acad. Sci. *96*, 2571–2573.

Melchers, F. (2015). Checkpoints that control B cell development. J. Clin. Invest. *125*, 2203–2210.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. Nat. Rev. Genet. *10*.

Mercolino, T., Locke, A., Afshari, A., Sasser, D., Travis, W., Arnold, L., and Haughton, G. (1989). Restricted immunoglobulin variable region gene usage by normal Ly-1 (CD5+) B cells that recognize phosphatidyl choline. J. Exp. Med. *169*, 1869–77.

Merrell, K.T., Benschop, R.J., Gauld, S.B., Aviszus, K., Decote-Ricardo, D., Wysocki, L.J., and Cambier, J.C. (2006). Identification of Anergic B Cells within a Wild-Type Repertoire. Immunity *25*.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet.

Mills, F.C., Harindranath, N., Mitchell, M., and Max, E.E. (1997). Enhancer Complexes Located Downstream of Both Human Immunoglobulin Cα Genes. J. Exp. Med. *186*, 845–858.

Minegishi, Y., and Conley, M.E. (2001). Negative selection at the pre-BCR checkpoint elicited by human mu heavy chains with unusual CDR3 regions. Immunity *14*, 631–41.

Minegishi, Y., E, C.-S., and Wang, Y. (1998). Mutations in the human λ5/14.1 gene result in B cell deficiency and agammaglobulinemia.

Mitchell, J.A., and Fraser, P. (2008). Transcription factories are nuclear subcompartments that remain in the absence of transcription. Genes Dev. *22*, 20–25.

Monaco, G., Chen, H., Poidinger, M., Chen, J., Magalhães, J.P. de, and Larbi, A. (2016). flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. Bioinforma. Oxf. Engl.

Mondon, P., Dubreuil, O., Bouayadi, K., and Kharrat, H. (2008). Human antibody libraries: a race to engineer and explore a larger diversity. Front. Biosci. J. Virtual Libr. *13*, 1117–1129.

Müeller-Storm, H.P., Sogo, J.M., and Schaffner, W. (1989). An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge. Cell *58*, 767–777.

Mundt, C., Licence, S., Shimizu, T., Melchers, F., and Mårtensson, I. (2001a). Loss of precursor B cell expansion but not allelic exclusion in VpreB1/VpreB2 double-deficient mice. J. Exp. Med. *193*, 435–45.

Mundt, C.A., Nicholson, I.C., Zou, X., Popov, A.V., Ayling, C., and Brüggemann, M. (2001b). Novel Control Motif Cluster in the IgH δ-γ3 Interval Exhibits B Cell-Specific Enhancer Function in Early Development. J. Immunol. *166*, 3315–3323.

Munshaw, S., and Kepler, T.B. (2010). SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements. Bioinformatics *26*, 867–872.

Murphy, A.J., Macdonald, L.E., Stevens, S., Karow, M., Dore, A.T., Pobursky, K., Huang, T.T., Poueymirou, W.T., Esau, L., Meola, M., et al. (2014). Mice with megabase humanization of their immunoglobulin genes generate antibodies as efficiently as normal mice. Proc. Natl. Acad. Sci. U. S. A. *111*, 5153–8.

Muyldermans, S. (2013). Nanobodies: natural single-domain antibodies. Annu. Rev. Biochem. *82*, 775–797.

Muyldermans, S., Atarhouch, T., Saldanha, J., Barbosa, J.A., and Hamers, R. (1994). Sequence and structure of VH domain from naturally occurring camel heavy chain immunoglobulins lacking light chains. Protein Eng. *7*, 1129–35.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

Nagata, K., Nakamura, T., Kitamura, F., Kuramochi, S., Taki, S., Campbell, K.S., and Karasuyama, H. (1997). The Igα/Igβ Heterodimer on μ-Negative ProB Cells Is Competent for Transducing Signals to Induce Early B Cell Differentiation. Immunity *7*, 559–570.

Nakamura, T., Okamoto, I., Sasaki, K., Yabuta, Y., Iwatani, C., Tsuchiya, H., Seita, Y., Nakamura, S., Yamamoto, T., and Saitou, M. (2016). A developmental coordinate of pluripotency among mice, monkeys and humans. Nature *537*, 57–62.

Nedbal, J., Hobson, P.S., Fear, D.J., Heintzmann, R., and Gould, H.J. (2012). Comprehensive FISH Probe Design Tool Applied to Imaging Human Immunoglobulin Class Switch Recombination. PLoS ONE *7*.

Nemazee, D. (2006). Receptor editing in lymphocyte development and central tolerance. Nat. Rev. Immunol. *6*, 728–740.

Nguyen, T.T., Elsner, R.A., and Baumgarth, N. (2015). Natural IgM prevents autoimmunity by enforcing B cell central tolerance induction. J. Immunol. Baltim. Md 1950 *194*, 1489–502.

Nguyen, T.T.T., Kläsener, K., Zürn, C., Castillo, P.A., Brust-Mascher, I., Imai, D.M., Bevins, C.L., Reardon, C., Reth, M., and Baumgarth, N. (2017). The IgM receptor FcμR limits tonic BCR signaling by regulating expression of the IgM BCR. Nat. Immunol. *18*, 321–333.

Nguyen, V.K., Hamers, R., Wyns, L., and Muyldermans, S. (2000). Camel heavy-chain antibodies: diverse germline VHH and specific mechanisms enlarge the antigen-binding repertoire. EMBO J. *19*, 921–930.

Nitschke, L., Kestler, J., Tallone, T., Pelkonen, S., and Pelkonen, J. (2001). Deletion of the DQ52 Element Within the Ig Heavy Chain Locus Leads to a Selective Reduction in VDJ Recombination and Altered D Gene Usage. J. Immunol. *166*, 2540–2552.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Berkum, N.L. van, Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Norton, H.K., Emerson, D.J., Huang, H., Kim, J., Titus, K.R., Gu, S., Bassett, D.S., and Phillips-Cremins, J.E. (2018). Detecting hierarchical genome folding with network modularity. Nat. Methods *15*, 119–122.

Oettinger, M.A., Schatz, D.G., Gorka, C., and Baltimore, D. (1990). RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science *248*, 1517–1523.

Ohnishi, K., and Melchers, F. (2003). The nonimmunoglobulin portion of λ5 mediates cell-autonomous pre-B cell receptor signaling. Nat. Immunol. *4*.

Olivares-Chauvet, P., Mukamel, Z., Lifshitz, A., Schwartzman, O., Elkayam, N., Lubling, Y., Deikus, G., Sebra, R.P., and Tanay, A. (2016). Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature *540*, 296–300.

Osborn, M.J., Ma, B., Avis, S., Binnie, A., Dilley, J., Yang, X., Lindquist, K., Ménoret, S., Iscache, A.-L., Ouisse, L.-H., et al. (2013). High-Affinity IgG Antibodies Develop Naturally in Ig-Knockout Rats Carrying Germline Human IgH/Igκ/Igλ Loci Bearing the Rat CH Region. J. Immunol. *190*, 1481–1490.

Osipovich, O.A., Subrahmanyam, R., Pierce, S., Sen, R., and Oltz, E.M. (2009). Cutting Edge: SWI/SNF Mediates Antisense Igh Transcription and Locus-Wide Accessibility in B Cell Precursors. J. Immunol. *183*, 1509–1513.

Ota, T., and Nei, M. (1994). Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol. Biol. Evol. *11*, 469–482.

Oudelaar, A.M., Davies, J.O.J., Downes, D.J., Higgs, D.R., and Hughes, J.R. (2017). Robust detection of chromosomal interactions from small numbers of cells using low-input Capture-C. Nucleic Acids Res. *45*, e184.

Papavasiliou, F., Misulovin, Z., Suh, H., and Nussenzweig, M. (1995). The role of Ig beta in precursor B cell transition and allelic exclusion. Science *268*, 408–11.

Parker, M.J., Licence, S., Erlandsson, L., Galler, G.R., Chakalova, L., Osborne, C.S., Morgan, G., Fraser, P., Jumaa, H., Winkler, T.H., et al. (2005). The pre-B-cell receptor induces silencing of VpreB and lambda5 transcription. EMBO J. *24*, 3895–905.

Pastor, W.A., Chen, D., Liu, W., Kim, R., Sahakyan, A., Lukianchikov, A., Plath, K., Jacobsen, S.E., and Clark, A.T. (2016). Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory. Cell Stem Cell *18*, 323–329.

Patel, R., Lin, M., Laney, M., Kurn, N., Rose, S., and Ullman, E. (1996). Formation of chimeric DNA primer extension products by template switching onto an annealed downstream oligonucleotide. Proc. Natl. Acad. Sci. U. S. A. *93*, 2969–74.

Paul, W.E. (2012). Fundamental Immunology (Philadelphia: Lippincott Williams and Wilkins).

Paulson, J.R., and Laemmli, U.K. (1977). The structure of histone-depleted metaphase chromosomes. Cell *12*, 817–828.

Pearson, J.C., Lemons, D., and McGinnis, W. (2005). Modulating Hox gene functions during animal body patterning. Nat. Rev. Genet. *6*, 893–904.

Pelanda, R., and Torres, R.M. (2012). Central B-Cell Tolerance: Where Selection Begins. Cold Spring Harb. Perspect. Biol. *4*, a007146.

Pelanda, R., Schwers, S., Sonoda, E., Torres, R.M., Nemazee, D., and Rajewsky, K. (1997). Receptor Editing in a Transgenic Mouse Model: Site, Efficiency, and Role in B Cell Tolerance and Antibody Diversification. Immunity *7*, 765–775.

Pérez, J.M.J., Renisio, J.G., Prompers, J.J., van Platerink, C.J., Cambillau, C., Darbon, H., and Frenken, L.G.J. (2001). Thermal Unfolding of a Llama Antibody Fragment: A Two-State Reversible Process†. Biochemistry (Mosc.) *40*, 74–83.

Perlot, T., Alt, F.W., Bassing, C.H., Suh, H., and Pinaud, E. (2005). Elucidation of IgH intronic enhancer functions via germ-line deletion. Proc. Natl. Acad. Sci. *102*, 14362–14367.

Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell *165*, 1012–1026.

Phillips, J.E., and Corces, V.G. (2009). CTCF: Master Weaver of the Genome. Cell *137*, 1194–1211.

Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M., and van Steensel, B. (2006). Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nat. Genet. *38*, 1005–1014.

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

Pillai, S., Cariappa, A., and Moran, S.T. (2004). Positive selection and lineage commitment during peripheral B-lymphocyte development. Immunol. Rev. *197*, 206–218.

Pillai, S., Cariappa, A., and Moran, S.T. (2005). MARGINAL ZONE B CELLS. Annu. Rev. Immunol. *23*.

Qiu, P., Simonds, E.F., Bendall, S.C., Jr, K.D.G., Bruggner, R.V., Linderman, M.D., Sachs, K., Nolan, G.P., and Plevritis, S.K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotechnol. *29*, 886–891.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Radic, M.Z., Mackle, J., Erikson, J., Mol, C., Anderson, W.F., and Weigert, M. (1993). Residues that mediate DNA binding of autoimmune antibodies. J. Immunol. Baltim. Md 1950 *150*, 4966–77.

Ragoczy, T., Bender, M.A., Telling, A., Byron, R., and Groudine, M. (2006). The locus control region is required for association of the murine β-globin locus with engaged transcription factories during erythroid maturation. Genes Dev. *20*.

Ralph, D.K., and Matsen, F.A. (2016). Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. PLOS Comput. Biol. *12*, e1004409.

Ramadani, F., Bolland, D.J., Garcon, F., Emery, J.L., Vanhaesebroeck, B., Corcoran, A.E., and Okkenhaug, K. (2010). The PI3K isoforms p110alpha and p110delta are essential for pre-B cell receptor signaling and B cell development. Sci. Signal. *3*.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Ramsden, D.A., Baetz, K., and Wu, G.E. (1994). Conservation of sequence in recombination signal sequence spacers. Nucleic Acids Res. *22*, 1785–1796.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–80.

Reddington, J.P., Perricone, S.M., Nestor, C.E., Reichmann, J., Youngson, N.A., Suzuki, M., Reinhardt, D., Dunican, D.S., Prendergast, J.G., Mjoseng, H., et al. (2013). Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. Genome Biol. *14*, R25.

Reichlin, A., Hu, Y., Meffre, E., Nagaoka, H., Gong, S., Kraus, M., Rajewsky, K., and Nussenzweig, M.C. (2001). B cell development is arrested at the immature B cell stage in mice carrying a mutation in the cytoplasmic domain of immunoglobulin beta. J. Exp. Med. *193*, 13–23.

Ren, L., Zou, X., Smith, J.A., and Brüggemann, M. (2004). Silencing of the immunoglobulin heavy chain locus by removal of all eight constant-region genes in a 200-kb region. Genomics *84*, 686–95.

Ren, W., Grimsholm, O., Bernardi, A.I., Höök, N., Stern, A., Cavallini, N., and Mårtensson, I.-L.L. (2015). Surrogate light chain is required for central and peripheral B-cell tolerance and inhibits anti-DNA antibody production by marginal zone B cells. Eur. J. Immunol. *45*, 1228–37.

Reynolds, A.E., Kuraoka, M., and Kelsoe, G. (2015). Natural IgM Is Produced by CD5− Plasma Cells That Occupy a Distinct Survival Niche in Bone Marrow. J. Immunol. *194*, 231–242.

Robinson, W.H. (2015). Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. Nat. Rev. Rheumatol. *11*, 171–182.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative Genomics Viewer. Nat. Biotechnol. *29*, 24–26.

Roldán, E., Fuxa, M., Chong, W., Martinez, D., Novatchkova, M., Busslinger, M., and Skok, J.A. (2005). Locus “decontraction” and centromeric recruitment contribute to allelic exclusion of the immunoglobulin heavy-chain gene. Nat. Immunol. *6*, 31–41.

Rolink, A., Grawunder, U., Winkler, T.H., Karasuyama, H., and Melchers, F. (1994). IL-2 receptor alpha chain (CD25, TAC) expression defines a crucial stage in pre-B cell development. Int. Immunol. *6*, 1257–64.

Rolink, A.G., Andersson, J., and Melchers, F. (1998). Characterization of immature B cells by a novel monoclonal antibody, by turnover and by mitogen reactivity. Eur. J. Immunol. *28*, 3738–3748.

Rolink, A.G., Winkler, T., Melchers, F., and Andersson, J. (2000). Precursor B cell receptor dependent B cell proliferation and differentiation does not require the bone marrow or fetal liver environment. J. Exp. Med. *191*, 23–32.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science *360*, 176–182.

Rouaud, P., Vincent-Fabert, C., Saintamand, A., Fiancette, R., Marquet, M., Robert, I., Reina-San-Martin, B., Pinaud, E., Cogné, M., and Denizot, Y. (2013). The IgH 3’ regulatory region controls somatic hypermutation in germinal center B cells. J. Exp. Med. *210*, 1501–1507.

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. Clifton NJ *132*, 365–386.

Rumfelt, L.L., Zhou, Y., Rowley, B.M., Shinton, S.A., and Hardy, R.R. (2006). Lineage specification and plasticity in CD19− early B cell precursors. J. Exp. Med. *203*, 675–687.

Sadlack, B., Löhler, J., Schorle, H., Klebb, G., Haber, H., Sickel, E., Noelle, R.J., and Horak, I. (1995). Generalized autoimmune disease in interleukin-2-deficient mice is triggered by an uncontrolled activation and proliferation of CD4+ T cells. Eur. J. Immunol. *25*, 3053–3059.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development *132*, 797–803.

Sahlén, P., Abdullayev, I., Ramsköld, D., Matskova, L., Rilakovic, N., Lötstedt, B., Albert, T.J., Lundeberg, J., and Sandberg, R. (2015). Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. Genome Biol. *16*, 156.

Saito, T., Chiba, S., Ichikawa, M., Kunisato, A., Asai, T., Shimizu, K., Yamaguchi, T., Yamamoto, G., Seo, S., Kumano, K., et al. (2003). Notch2 Is Preferentially Expressed in Mature B Cells and Indispensable for Marginal Zone B Lineage Development. Immunity *18*, 675–685.

Sambrook, J., and Russell, D. (2000). Molecular Cloning (3-volume set): A Laboratory Manual (Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press,U.S.).

Sanchez, P., Crain-Denoyelle, A.-M., Daras, P., Gendron, M.-C., and Kanellopoulos-Langevin, C. (2000). The level of expression of μ heavy chain modifies the composition of peripheral B cell subpopulations. Int. Immunol. *12*, 1459–1466.

Sano, K., Tanihara, H., Heimark, R.L., Obata, S., Davidson, M., St John, T., Taketani, S., and Suzuki, S. (1993). Protocadherins: a large family of cadherin-related molecules in central nervous system. EMBO J. *12*, 2249–2256.

Savage, H.P., and Baumgarth, N. (2015). Characteristics of natural antibody–secreting cells. Ann. N. Y. Acad. Sci. *1362*, 132–142.

Schatz, D.G., and Ji, Y. (2011). Recombination centres and the orchestration of V(D)J recombination. Nat. Rev. Immunol. *11*, 251–263.

Schatz, D.G., and Swanson, P.C. (2011). V(D)J recombination: mechanisms of initiation. Annu. Rev. Genet. *45*, 167–202.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M.M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. *25*, 582–97.

Schram, B.R., Tze, L.E., Ramsey, L.B., Liu, J., Najera, L., Vegoe, A.L., Hardy, R.R., Hippen, K.L., Farrar, M.A., and Behrens, T.W. (2008). B Cell Receptor Basal Signaling Regulates Antigen-Induced Ig Light Chain Rearrangements. J. Immunol. *180*, 4728–4741.

Schroeder, K., Herrmann, M., and Winkler, T.H. (2012). The role of somatic hypermutation in the generation of pathogenic antibodies in SLE. Autoimmunity *46*, 121–127.

Schuettengruber, B., Bourbon, H.-M., Di Croce, L., and Cavalli, G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell *171*, 34–57.

Schuh, W., Meister, S., Roth, E., and Jäck, H.-M. (2003). Cutting Edge: Signaling and Cell Surface Expression of a μH Chain in the Absence of λ5: A Paradigm Revisited. J. Immunol. *171*, 3343–3347.

Schultz, M., Clarke, S.H., Arnold, L.W., Sartor, R.B., and Tonkonogy, S.L. (2001). Disrupted B-lymphocyte development and survival in interleukin-2-deficient mice. Immunology *104*, 127–134.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Shaffer, A.L., and Schlissel, M.S. (1997). A truncated heavy chain protein relieves the requirement for surrogate light chains in early B cell development. J. Immunol. *159*, 1265–1275.

Shimizu, T., Mundt, C., Licence, S., Melchers, F., and Mårtensson, I.-L.L. (2002). VpreB1/VpreB2/lambda 5 triple-deficient mice show impaired B cell development but functional allelic exclusion of the IgH locus. J. Immunol. Baltim. Md 1950 *168*, 6286–93.

Shlomchik, M.J., and Weisel, F. (2012). Germinal center selection and the development of memory B and plasma cells. Immunol. Rev. *247*, 52–63.

Shugay, M., Britanova, O.V., Merzlyak, E.M., Turchaninova, M.A., Mamedov, I.Z., Tuganbaev, T.R., Bolotin, D.A., Staroverov, D.B., Putintseva, E.V., Plevova, K., et al. (2014). Towards error-free profiling of immune repertoires. Nat. Methods *11*, 653–5.

Silva, N.S.D., and Klein, U. (2015). Dynamics of B cells in germinal centres. Nat. Rev. Immunol. *15*, 137–48.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., Wit, E. de, Steensel, B. van, and Laat, W. de (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nat. Genet. *38*, 1348–1354.

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science *228*, 1315–1317.

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499.

Song, J., Uyttersprot, N., Classen, S., and Waisman, A. (2016). The IgG1 B-cell receptor provides survival and proliferative signals analogue to the Igα but not the Igβ co-receptor. Eur. J. Immunol. *46*, 1878–1886.

Spitz, F. (2016). Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. Semin. Cell Dev. Biol. *57*, 57–67.

Staudt, L.M., and Lenardo, M.J. (1991). Immunoglobulin Gene Transcription. Annu. Rev. Immunol. *9*, 373–398.

Steels, A., and Gettemans, L.B. and J. (2018). Use, Applications and Mechanisms of Intracellular Actions of Camelid VHHs. Antib. Eng.

Stirparo, G.G., Boroviak, T., Guo, G., Nichols, J., Smith, A., and Bertone, P. (2018). Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human pre-implantation epiblast. Development *145*, dev158501.

Stoops, J., Byrd, S., and Hasegawa, H. (2012). Russell body inducing threshold depends on the variable domain sequences of individual human IgG clones and the cellular protein homeostasis. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1823*, 1643–1657.

Stubbington, M.J., and Corcoran, A.E. (2013). Non-coding transcription and large-scale nuclear organisation of immunoglobulin recombination. Curr. Opin. Genet. Dev. *23*, 81–88.

Su, Y.-W.W., Flemming, A., Wossning, T., Hobeika, E., Reth, M., and Jumaa, H. (2003). Identification of a pre-BCR lacking surrogate light chain. J. Exp. Med. *198*, 1699–706.

Subrahmanyam, R., and Sen, R. (2010). RAGs' eye view of the immunoglobulin heavy chain gene locus. Semin. Immunol. *22*, 337–345.

Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. Nat. Protoc. *13*, 599–604.

Symmons, O., Uslu, V.V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. Genome Res. *24*, 390–400.

Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W., et al. (2014). Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. Cell *158*, 1254–1269.

Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D.G. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196–199.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L., et al. (2014). Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. Cell Stem Cell *15*, 471–487.

Theunissen, T.W., Friedli, M., He, Y., Planet, E., O'Neil, R.C., Markoulaki, S., Pontis, J., Wang, H., Iouranova, A., Imbeault, M., et al. (2016). Molecular Criteria for Defining the Naive Human Pluripotent State. Cell Stem Cell *19*, 502–515.

Thompson, E.C., Cobb, B.S., Sabbattini, P., Meixlsperger, S., Parelho, V., Liberg, D., Taylor, B., Dillon, N., Georgopoulos, K., Jumaa, H., et al. (2007). Ikaros DNA-binding proteins as integral components of B cell developmental-stage-specific regulatory circuits. Immunity *26*, 335–344.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic Stem Cell Lines Derived from Human Blastocysts. Science *282*, 1145–1147.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Tsiantoulas, D., Kiss, M., Bartolini-Gritti, B., Bergthaler, A., Mallat, Z., Jumaa, H., and Binder, C.J. (2017). Secreted IgM deficiency leads to increased BCR signaling that results in abnormal splenic B cell development. Sci. Rep. *7*, 3540.

Turchaninova, M., Davydov, A., Britanova, O., Shugay, M., Bikos, V., Egorov, E., Kirgizova, V., Merzlyak, E., Staroverov, D., Bolotin, D., et al. (2016). High-quality full-length immunoglobulin profiling with unique molecular barcoding. Nat. Protoc. *11*, 1599–1616.

Turinetto, V., and Giachino, C. (2015). Histone variants as emerging regulators of embryonic stem cell identity. Epigenetics *10*, 563–573.

Tze, L.E., Schram, B.R., Lam, K.-P.P., Hogquist, K.A., Hippen, K.L., Liu, J., Shinton, S.A., Otipoby, K.L., Rodine, P.R., Vegoe, A.L., et al. (2005). Basal immunoglobulin signaling actively maintains developmental stage in immature B cells. PLoS Biol. *3*, e82.

Valetti, C., Grossi, C.E., Milstein, C., and Sitia, R. (1991). Russell bodies: a general response of secretory cells to synthesis of a mutant immunoglobulin which can neither exit from, nor be degraded in, the endoplasmic reticulum. J. Cell Biol. *115*, 983–94.

Vallot, C., Patrat, C., Collier, A.J., Huret, C., Casanova, M., Liyakat Ali, T.M., Tosolini, M., Frydman, N., Heard, E., Rugg-Gunn, P.J., et al. (2017). XACT Noncoding RNA Competes with XIST in the Control of X Chromosome Activity during Human Early Development. Cell Stem Cell *20*, 102–111.

Vanhove, M., Usherwood, Y.-K., and Hendershot, L.M. (2001). Unassembled Ig Heavy Chains Do Not Cycle from BiP In Vivo but Require Light Chains to Trigger Their Release. Immunity *15*, 105–114.

Varriale, S., Merlino, A., Coscia, M., Mazzarella, L., and Oreste, U. (2010). An evolutionary conserved motif is responsible for Immunoglobulin heavy chain packing in the B cell membrane. Mol. Phylogenet. Evol. *57*, 1238–44.

Verkoczy, L., Duong, B., Skog, P., Aït-Azzouzene, D., Puri, K., Vela, J.L., and Nemazee, D. (2007). Basal B Cell Receptor-Directed Phosphatidylinositol 3-Kinase Signaling Turns Off RAGs and Promotes B Cell-Positive Selection. J. Immunol. *178*, 6332–6341.

Victora, G.D., and Nussenzweig, M.C. (2012). Germinal Centers. Annu. Rev. Immunol. *30*, 429–457.

Vincent-Fabert, C., Fiancette, R., Pinaud, E., Truffinet, V., Cogné, N., Cogné, M., and Denizot, Y. (2010). Genomic deletion of the whole IgH 3′ regulatory region (hs3a, hs1,2, hs3b, and hs4) dramatically affects class switch recombination and Ig secretion to all isotypes. Blood *116*, 1895–1898.

Volpi, S.A., Verma-Gaur, J., Hassan, R., Ju, Z., Roa, S., Chatterjee, S., Werling, U., Hou, H., Will, B., Steidl, U., et al. (2012). Germline Deletion of Igh 3′ Regulatory Region Elements hs 5, 6, 7 (hs5–7) Affects B Cell-Specific Regulation, Rearrangement, and Insulation of the Igh Locus. J. Immunol. *188*, 2556–2566.

Vree, P.J. de, Wit, E. de, Yilmaz, M., Heijning, M. van de, Klous, P., Verstegen, M.J., Wan, Y., Teunissen, H., Krijger, P.H., Geeven, G., et al. (2014). Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. Nat. Biotechnol. *32*, 1019–25.

Waisman, A., Kraus, M., Seagal, J., Ghosh, S., Melamed, D., Song, J., Sasaki, Y., Classen, S., Lutz, C., Brombacher, F., et al. (2007). IgG1 B cell receptor signaling is inhibited by CD22 and promotes the development of B cells whose survival is less dependent on Igα/β. J. Exp. Med. *204*, 747–758.

Wakabayashi, C., Adachi, T., Wienands, J., and Tsubata, T. (2002). A distinct signaling pathway used by the IgG-containing B cell antigen receptor. Science *298*, 2392–2395.

Wang, B., DeKosky, B.J., Timm, M.R., Lee, J., Normandin, E., Misasi, J., Kong, R., McDaniel, J.R., Delidakis, G., Leigh, K.E., et al. (2018). Functional interrogation and mining of natively paired human V_H:V_L antibody repertoires. Nat. Biotechnol. *36*, 152–155.

Wang, H., Coligan, J.E., and Morse, H.C.I. (2016). Emerging Functions of Natural IgM and Its Fc Receptor FCMR in Immune Homeostasis. Front. Immunol. *7*.

Wang, J., Lawry, S.T., Cohen, A.L., and Jia, S. (2014). Chromosome boundary elements and regulation of heterochromatin spreading. Cell. Mol. Life Sci. CMLS *71*, 4841–4852.

Ward, E.S., Güssow, D., Griffiths, A.D., Jones, P.T., and Winter, G. (1989). Binding activities of a repertoire of single immunoglobulin variable domains secreted from Escherichia coli. Nature *341*, 544–546.

Wardemann, H., and Busse, C.E. (2017). Novel Approaches to Analyze Immunoglobulin Repertoires. Trends Immunol. *38*, 471–482.

Wardemann, H., Yurasov, S., Schaefer, A., Young, J.W., Meffre, E., and Nussenzweig, M.C. (2003). Predominant autoantibody production by early human B cell precursors. Science *301*, 1374–7.

Weinstein, J.A., Jiang, N., White, R.A., Fisher, D.S., and Quake, S.R. (2009). High-Throughput Sequencing of the Zebrafish Antibody Repertoire. Science *324*, 807–810.

Wen, L., Brill-Dashoff, J., Shinton, S.A., Asano, M., Hardy, R.R., and Hayakawa, K. (2005). Evidence of Marginal-Zone B Cell- Positive Selection in Spleen. Immunity *23*, 297–308.

Werner, M., and Jumaa, H. (2015). Chapter 6 - Proliferation and Differentiation Programs of Developing B Cells. In Molecular Biology of B Cells (Second Edition), F.W. Alt, T. Honjo, A. Radbruch, and M. Reth, eds. (London: Academic Press), pp. 75–97.

Wijchers, P.J., and de Laat, W. (2011). Genome organization influences partner selection for chromosomal rearrangements. Trends Genet. *27*, 63–71.

Willerford, D.M., Chen, J., Ferry, J.A., Davidson, L., Ma, A., and Alt, F.W. (1995). Interleukin-2 receptor alpha chain regulates the size and content of the peripheral lymphoid compartment. Immunity *3*, 521–30.

Wilson, K.L., and Foisner, R. (2010). Lamin-binding Proteins. Cold Spring Harb. Perspect. Biol. *2*, a000554.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., and Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data. F1000Research.

Winkler, T.H., and Mårtensson, I.-L. (2018). The Role of the Pre-B Cell Receptor in B Cell Development, Repertoire Selection, and Tolerance. Front. Immunol. *9*.

Wit, E. de, and Laat, W. de (2012). A decade of 3C technologies: insights into nuclear organization. Genes Dev. *26*, 11–24.

Wit, E. de, Vos, E.S., Holwerda, S.J., Valdes-Quezada, C., Verstegen, M.J., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H., and Laat, W. de (2015). CTCF Binding Polarity Determines Chromatin Looping. Mol. Cell *60*, 676–84.

Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nat. Rev. Genet. *18*, 517–534.

Wu, T.T., Johnson, G., and Kabat, E.A. (1993). Length distribution of CDRH3 in antibodies. Proteins *16*, 1–7.

Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D.A., Stocsits, R.R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M.J., et al. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. EMBO J. *36*, 3573–3599.

Yaari, G., and Kleinstein, S.H. (2015). Practical guidelines for B-cell receptor repertoire sequencing analysis. *7*, 121.

Yamada, T. (2011). Therapeutic Monoclonal Antibodies. Keio J. Med. *60*, 37–46.

Yan, K.-K., Lou, S., and Gerstein, M. (2017). MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. PLOS Comput. Biol. *13*, e1005647.

Yang, Q., Riblet, R., and Schildkraut, C.L. (2005). Sites That Direct Nuclear Compartmentalization Are near the 5′ End of the Mouse Immunoglobulin Heavy-Chain Locus. Mol. Cell. Biol. *25*, 6021–6030.

Ye, J. (2004). The immunoglobulin IGHD gene locus in C57BL/6 mice. Immunogenetics *56*, 399–404.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics *13*, 134.

Ye, J., Ma, N., Madden, T.L., and Ostell, J.M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. *41*, W34–W40.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Zachau, H.G. (1993). The immunoglobulin κ locus — or — what has been learned from looking closely at one-tenth of a percent of the human genome. Gene *135*, 167–173.

Zemlin, M., Klinger, M., Link, J., Zemlin, C., Bauer, K., Engler, J.A., Schroeder, H.W., and Kirkham, P.M. (2003). Expressed Murine and Human CDR-H3 Intervals of Equal Length Exhibit Distinct Repertoires that Differ in their Amino Acid Composition and Predicted Range of Structures. J. Mol. Biol. *334*, 733–749.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet. *38*, 1341–1347.

Zheng, T., Hou, Y., Zhang, P., Zhang, Z., Xu, Y., Zhang, L., Niu, L., Yang, Y., Liang, D., Yi, F., et al. (2017). Profiling single-guide RNA specificity reveals a mismatch sensitive core sequence. Sci. Rep. *7*.

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E.A., Koebbe, B.C., Nielsen, C., Hirst, M., Farnham, P., et al. (2011). The Human Epigenome Browser at Washington University. Nat. Methods *8*, 989–990.

Zhou, X., Lowdon, R.F., Li, D., Lawson, H.A., Madden, P.A., Costello, J.F., and Wang, T. (2013). Exploring long-range genome interactions using the WashU Epigenome Browser. Nat. Methods *10*, 375–6.

Zhou, X., Li, D., Zhang, B., Lowdon, R.F., Rockweiler, N.B., Sears, R.L., Madden, P.A., Smirnov, I., Costello, J.F., and Wang, T. (2015). Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. Nat. Biotechnol. *33*, 345–346.

Zou, X., Xian, J., Popov, A.V., Rosewell, I.R., Müller, M., and Brüggemann, M. (1995). Subtle differences in antibody responses and hypermutation of lambda light chains in mice with a disrupted chi constant region. Eur. J. Immunol. *25*, 2154–62.

Zou, X., Piper, T.A., Smith, J.A., Allen, N.D., Xian, J., and Brüggemann, M. (2003). Block in Development at the Pre-B-II to Immature B Cell Stage in Mice Without Igκ and Igλ Light Chain. J. Immunol. *170*, 1354–1361.

Zou, X., Smith, J.A., Nguyen, V.K., Ren, L., Luyten, K., Muyldermans, S., and Brüggemann, M. (2005). Expression of a Dromedary Heavy Chain-Only Antibody and B Cell Development in the Mouse. J. Immunol. *175*, 3769–3779.

Zou, X., Osborn, M.J., Bolland, D.J., Smith, J.A., Corcos, D., Hamon, M., Oxley, D., Hutchings, A., Morgan, G., Santos, F., et al. (2007). Heavy chain–only antibodies are spontaneously produced in light chain–deficient mice. J. Exp. Med. *204*, 3271–3283.

Zou, X., Smith, J.A., Corcos, D., Matheson, L.S., Osborn, M.J., and Brüggemann, M. (2008). Removal of the BiP-retention domain in Cmicro permits surface deposition and developmental progression without L-chain. Mol. Immunol. *45*, 3573–9.

7 Appendix A

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material

The following text has been redacted due to sensitivity of the material