

Characterization of the Naïve Kappa Light Chain Murine  
Immunoglobulin Repertoire in Spaceflight

by

Claire Ward

B.A., MidAmerica Nazarene University, 2013

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Division of Biology  
College of Arts and Sciences

KANSAS STATE UNIVERSITY  
Manhattan, Kansas

2017

Approved by:

Major Professor  
Dr. Stephen Keith Chapes

## **Abstract**

Immunoglobulins are receptors expressed on the outside of a B cell that can specifically bind pathogens and toxic substances within a host. These receptors are heterodimers of two chains: heavy and light, which are encoded at separate loci. Enzymatic splicing of gene segments at heavy and light chain loci within the genomic DNA in every B cell results in a highly diversified and specific repertoire of immunoglobulins in a single host. Spaceflight is known to affect reduce splenic B cell populations and B cell progenitors within the bone marrow, potentially restricting the diversity of the immunoglobulin repertoire (Ig-Rep).

The objective of this thesis project was to characterize the impact of spaceflight on the kappa light-chain Ig-Rep of the C57BL/6 mouse. High-throughput sequencing (HTS) technologies have enabled the rapid characterization of Ig-Reps, however, standard Ig-Rep workflows often rely the amplification of immunoglobulin sequences to ensure the capture immunoglobulin sequences from rare B cell clones. Additionally, the Ig-Rep is often assessed in sorted B cell populations.

Opportunities for spaceflight experiments are limited and costly, and the exclusive amplification of immunoglobulin sequences prior to HTS results in a dataset that cannot be mined for additional information. Furthermore, due to the difficulties of tissue collection in spaceflight, HTS of sorted B cell populations is not feasible. We optimized a protocol in which the Ig-Rep was assessed from unamplified whole tissue immunoglobulin transcripts. The Ig-Rep was characterized by gene segment usage, gene segment combinations and the region in which gene segments are joined. HTS datasets of ground control animals and animals flown aboard the International Space Station were compared to explore the impact of spaceflight on the unimmunized murine Ig-Rep.

# Table of Contents

List of Figures .....	v
List of Tables .....	vi
List of Appendices .....	vii
Acknowledgements .....	viii
Chapter 1 - Introduction .....	1
Generation of the Antibody Repertoire .....	1
High-throughput Sequencing and the Antibody Repertoire .....	9
Spaceflight and the Adaptive Immune Response .....	19
Objective .....	24
References .....	25
Chapter 2 - Validation of Methods to Assess the Immunoglobulin Gene Repertoire in Tissues	
Obtained from Mice on the International Space Station .....	60
Abstract .....	60
Introduction .....	61
Materials and Methods .....	64
Results .....	69
Discussion .....	76
Figures and Tables .....	83
References .....	93
Chapter 3 - Characterization of the Naïve Murine Antibody Repertoire Using Unamplified High-throughput Sequencing .....	102
Abstract .....	102
Introduction .....	103
Materials and Methods .....	104
Results .....	108
Discussion .....	116
Figures and Tables .....	122
References .....	132

Chapter 4 - Effects of Spaceflight on the Antibody Repertoire of Unimmunized C57BL/6 Mice	141
.....	141
Abstract.....	141
Abstract.....	141
Introduction.....	142
Materials and Methods.....	144
Results.....	147
Discussion.....	159
Figures and Tables.....	165
References.....	180
Chapter 5 - Conclusions.....	193

## List of Figures

Figure 2.1 Bioinformatic analysis workflows.....	83
Figure 2.2 Decision-making matrix to remove duplicate reads after IMGT processing .....	84
Figure 2.3 Top ten VH gene segments used among treatment groups .....	85
Figure 2.4 Top ten V $\kappa$ used among treatment groups.....	86
Figure 2.5 D, J, and heavy chain constant usage among treatment groups .....	87
Figure 2.6 CDR3 AA sequence usage among treatment groups .....	88
Figure 2.7 Correlation of V gene segments between genome and reference mapping.....	89
Figure 3.1 V-gene segment usage among unimmunized mouse poo.....	122
Figure 3.2 V-gene segment usage by chromosomal location .....	123
Figure 3.3 D-, J-gene segment and constant region usage.....	124
Figure 3.4 V(D)J-gene segment combinations .....	125
Figure 3.5 CDR3 length.....	127
Figure 3.6 Overlap of unique CDR3 sequences within pools and top CDR3 sequences .....	128
Figure 3.7 Alignments of top V(D)J-gene segment combinations .....	129
Figure 4.1 Expression of top V-gene segments .....	165
Figure 4.2 Expression of top V $\kappa$ -gene segments from spleen and liver .....	166
Figure 4.3 Expression of DH- and J-gene segments and IgH constant region usage .....	167
Figure 4.4 Gene segment combinations in ground control and flight animals .....	168
Figure 4.5 CDR3 length in IgH and Igk sequences .....	170
Figure 4.6 Top CDR3 usage and overlap of CDR3 between treatment animals .....	171
Figure 4.7 Nucleotide alignment of CDR3 from top V-D-J combination .....	172
Figure 4.8 Substitution mutations by Ig region .....	174

## List of Tables

Table 2.1 Sequences used for heavy chain identification.....	90
Table 2.2 Sequencing and mapping results from the cells, tissue, and size selected treatment groups.....	91
Table 2.3 Comparison of mapping techniques in HiSeq datasets.....	92
Table 3.1 Sequencing and mapping statistics from mouse pools 1, 2, and 3.....	131
Table 4.1 Spleen sequencing read counts in ground (G) and flight (F) mice.....	175
Table 4.2 Comparison of flight and ground V-gene segment usage.....	176
Table 4.3 V $\kappa$ liver sequencing read counts in ground (G) and flight (F) mice.....	177
Table 4.4 V-J linear regression analyses.....	178
Table 4.5 CDR3 length by isotype.....	179

## List of Appendices

Appendix A.1 Statement of copyright release .....	196
Appendix A.2 V-gene segment usage in normal mouse pools .....	197
Appendix A.3 CDR3 length by individual mouse pool .....	199
Appendix A.4 CDR3 sequences shared among all pools .....	200
Appendix A.5 Alignment of top IgH gene segment combination .....	205
Appendix A.6 V-gene segment heat maps.....	207
Appendix A.7 D- and J-gene segment and constant region heat maps.....	212
Appendix A.8 V/J combinations of individual animals.....	213
Appendix A.9 Top V(D)J gene family combinations.....	215
Appendix B.1 Immunoglobulin reference sequences .....	217
Appendix B.2 Mapping.....	220
Appendix B.3 IMGT submission.....	227
Appendix B.4 Cleaning IMGT output data.....	229
Appendix B.5 Duplicate read removal.....	237
Appendix B.6 Assessment of V Gene Segment Usage.....	242
Appendix B.7 J gene segment usage.....	249
Appendix B.8 Gene segment combination .....	252
Appendix B.9 CDR3 analyses .....	258

## **Acknowledgements**

I would like to express my sincere appreciation to my committee members: Dr. Stephen Keith Chapes, Dr. Sherry Fleming, and Dr. Carol Chitko-McKown. Thank you for your guidance on my project. I am thankful for this experience and I am excited to continue to build on what I have learned from you at Kansas State University.

I would also like to thank members of the Chapes lab: Trisha Rettig, Bailey Bye, and Savannah Hlavachek. It has been a privilege to work as a team on this project. Finally, I would like to thank husband and my family for all of their support.

This work has been supported by NASA grants NNX13AN34G and NNX15AB45G, NIH grant GM103418, the Molecular Biology Core supported by the College of Veterinary Medicine at Kansas State University, and the Kansas State University Terry C. Johnson Center for Basic Cancer Research.



# **Chapter 1 - Introduction**

## **Generation of the Antibody Repertoire**

### **Early Immunology and the Discovery of the B Cell**

Just before the turn of the twentieth century, Emil von Behring and Shibasaburo Kitasato showed that sera from the blood of animals immunized against diphtheria and tetanus had the ability protect other animals both before and after challenge with live bacteria or purified bacteria toxin (1, 2). Paul Ehrlich systematically characterized the action of these “antitoxins”, which we now know to be immunoglobulins (Ig). Interestingly, Ehrlich remarked at the improbability these molecules being designed specifically for the purpose of neutralizing toxins, especially considering that the introduction of such toxins was often “only arbitrarily brought into relation with them by the will of the investigator” (3). Ehrlich went on to propose that these antitoxins existed as side chains on cells that served other purposes such as nutrient uptake, interacting with either nutrient or toxin like a lock and key, that could be shed from the cell (3).

Indeed, immunoglobulins were found to be membrane bound, cellular proteins, which could be secreted as antibodies upon immunization and subsequent antigen challenge (4, 5). Antibody production originates in white blood cells called B lymphocytes (6), originally named for their discovery within an avian organ, the bursa of Fabricius (7). In mammals, B cells originate from hematopoietic stem cells within the bone marrow and fetal liver. In the late 1950s, Nossal and Lederberg determined that antibody producing cells were capable of producing only one type of antibody through single cell analysis of lymphatic tissue from mice that were immunized and challenged with one of two staphylococcal strains, also demonstrating a high degree of specificity in antibodies (5). This study provided evidence for clonal selection theory, in which antibody producing cells produce only one type of antibody, whereas Ehrlich’s side chain hypotheses

suggested the presence of many different antibodies on a cell's surface. Thus, the collection of single immunoglobulin types expressed on B cells within an individual is the immunoglobulin repertoire from which the individual combats antigenic challenges.

### **Structural Studies of Immunoglobulins**

Ehrlich's early skepticism that antibodies existed singly and specifically for virtually any challenge, no matter how unnatural, was understandable through the lens of the "one-enzyme, one-gene" hypothesis, in which a single gene was understood to be responsible for one protein product that had a specific function within the cell. A one-to-one ratio of gene to antibody would presumably amount to an unrealistically large portion of the genome, an issue that would later be resolved through structural and genetic studies.

In the late 1950s enzymatic digests and reducing agents were paired with column filtration, ultracentrifugation, and size determination by gel electrophoresis to assess antibody structure. A papain digestion of antibodies yielded three fragments that occurred at a 1:1:1 ratio (8). Two of the fragments were found to have similar molecular weights (50-53 kDa) and amino acid composition. These fragments were determined to be identical and noted to be capable of antigen binding and were thus termed "Fab". The third fragment, termed "fragment crystallizable" (Fc), was not capable of binding to antigen and had an amino acid composition that was different from Fab fragments. Similarly, pepsin digestion of antibodies that resulted in three fragments similar to those resulting from papain digestion, although the cleavage site differed slightly (9). The two univalent Fab fragments are joined with the Fc fragment by flexible hinge regions, forming a bivalent structure (10). Antibody structure was further characterized through reduction of disulfide bonds present within the antibody. The reduced products weighed between 30 and 50 kDa which

later became known as light and heavy chains, respectively (11). Comparison of antibody digests for specific antigens revealed small fragments that varied by antibody specificity and large fragments that remained similar (11).

The observation that antibodies consist of constant and variable regions was further supported by comparing the amino acid sequences of light-chains found in the urine of B-cell myeloma patients, known as Bence-Jones proteins (12-14). Assessment of conserved amino acid sequences among Bence-Jones proteins revealed that light-chains existed in two classes, kappa and lambda (15). Heavy chains were also found to exist in different classes that correspond to the constant region amino acid sequence being used: IgM ( $C_\mu$ ), IgD ( $C_\delta$ ), IgE ( $C_\epsilon$ ), IgA ( $C_\alpha$ ) and IgG ( $C_\gamma$ ), the latter two of which can be subdivided into multiple subclasses in both humans and mice (16). Overall, these studies revealed that antibodies were heterodimers of heavy and light chains which each contained conserved and variable regions. Furthermore, the data suggested that it was likely that three genetic loci existed: a heavy chain locus, a kappa light chain locus and a lambda light chain locus.

### **Antibody Diversification is Achieved through Somatic Rearrangement of Ig Loci**

Indeed, the heavy- (IgH) and light-immunoglobulin (IgL) chains were encoded in separate loci. In mice, the kappa (Ig $\kappa$ ) and lambda (Ig $\lambda$ ) chain loci are found on chromosomes 6 and 16, respectively while the IgH locus is found on chromosome 12 (17-19). Due to the presence of conserved and variable region domains, Dreyer and Bennet proposed that immunoglobulin genes were able to undergo a rearrangement in early B cell development (20). Evidence for rearrangement, known as somatic recombination, came from studies comparing immunoglobulin sequences of murine myeloma cell lines with immunoglobulin sequences from mouse embryos

(21-27). Heavy- and light-chain sequences isolated from the myelomas was found to be much smaller than the IgH and IgL loci found within embryonic cells. Recombination was ultimately the result of not only the combination of variable and constant regions, but also the combinations of gene segments within the variable region.

The Ig $\kappa$  locus consists of a single constant region and a variable region that contains multiple gene segments (23, 26). The variable- (V) gene segment is roughly 100 amino acids in length (23) and the current mouse reference genome contains 151 segments that can be expressed in a functional antibody (23) (NCBI: GRCm38. NC\_000072.6). There are four shorter (roughly 10 amino acids) gene segments that join the variable gene segment to the constant region, or joining- (J) gene segments, encoded within the current mouse reference genome (26). Similarly, the Ig $\lambda$  locus was found to consist of V- and J-gene segments, however, the Ig $\lambda$  locus only encodes 3 V-gene segments within the mouse genome and has 4 J-gene segments downstream of the variable gene segments which are positioned next to and pair directly with their own constant regions (25, 28, 29) (NCBI: GRCm38.p4 NC\_000082.6). Although IgH contains multiple constant regions as seen with Ig $\lambda$ , the IgH locus is arranged more similarly to the Ig $\kappa$  locus in that V-gene segments are located upstream of all constant regions (27, 30, 31) (NCBI: GRCm38.p4 NC\_000078.6). IgH contains an additional small gene segment that results in more combinatorial antibody diversity (D), which is located between V- and J-gene segments (30).

Within the variable region there are three hypervariable regions, flanked on either side by regions that are less variable, known as framework regions. These hypervariable regions are known as complementarity determining regions as they are the largest contributors to the antigen specificity, or idiootype, of the immunoglobulin (32, 33). CDR1 and 2 are contained within the V-gene segment, whereas CDR3 is located in the region in which V- and J-gene segments combine

in IgL and V-, D-, and J-gene segments combined in IgH (34). The joining of these gene segments and combination of IgH and IgL constitute two major mechanisms for antibody diversity, which is not limited to combinatorial diversity alone upon further characterization somatic V(D)J recombination.

### **Mechanisms Behind the Generation of Antibody Diversity**

Somatic recombination of V(D)J-gene segments is facilitated by a number of enzymes and binding proteins that are either specific to V(D)J recombination or also involved in DNA double-strand break repair (DSBR). The first step in recombination is the recognition of recombination signal sequences (RSS) that consist of conserved heptameric and nonameric nucleotide sequences that are separated by 12- or 23-basepair spacers (34-36). RSS are found downstream of V-gene segments, flank both sides of D-gene segments, and are upstream of J-gene segments (34-36). Pairing of multiple V-gene segments to one another is prevented through the joining of gene segments with a 12-bp spacer to only a gene segment that has a 23-bp spacer; this is known as the 12/23 pairing rule (37). For instance, Ig $\kappa$  V-gene segments, which are followed by a nonmer-12-bp-heptamer RSS, can only be paired with a J-gene segment, which has a heptamer-23-bp-nonamer RSS (35). These RSS are bound by the enzymes encoded by *recombination activation genes 1* and *2* (RAG1/2) (38, 39) and high mobility group proteins 1 and 2 (HMG1/2), which stabilize the binding of RAG1/2 to RSS (40, 41). Rag1/2 creates a single strand nick between signal and coding sequences resulting in a double-stranded break, that yields sealed hairpins at the coding joints and blunt ends at the signal joint (42). The signal joint ends are held in synaptic complex and bound by protein kinases consisting of 70 and 80 kDa Ku subunits which recruit DNA ligase IV (43, 44).

The protein XRCC4 then associates with DNA ligase IV, enabling ligation of the signal joint into a non-coding extrachromosomal loop (45, 46).

In contrast to the direct and precise joining of the signal ends, additional variability is introduced through the addition of DNA templated and non-templated nucleotides to create an imprecise coding joint (45). The hairpins of the coding gene segments are bound by 70 and 80 kDa Ku protein kinase subunits and are randomly cleaved by the enzyme Artemis (47). Cleavage of the hairpin in a position that generates single-stranded DNA of previously base-paired nucleotides results in the templated introduction of palindromic (p-) nucleotides (48). The cleaved ends are then bound by Terminal Deoxynucleotidyl Transferase (TdT) which can add non-templated (n-) nucleotides (49, 50). Finally, the DNA ends are ligated by DNA ligase IV and XRCC4 to yield a coding joint as described above with signal joints (46). Mutations of proteins involved in the somatic recombination process result prevent V(D)J recombination, which can result in severe combined immunodeficiency (SCID), and may also increase risk for cancer from ionizing radiation due to the overlap of BSBR machinery (51-55).

### **Timing of Immunoglobulin Development**

In mammals, B cells first arise from hematopoietic stem cells within the fetal liver during embryonic development and within the bone marrow in adults. In both the fetal liver and bone marrow, B-cell lineage is divided into several phases. B-cell lineage commitment of stem cells through the expression of CD45<sup>+</sup> results in initial expression TdT and RAG1/2 in the pro-B cell (56, 57). It has been shown that within the fetal liver, expression of TdT is low, resulting in fewer addition of n-nucleotides (56, 58, 59). In the pro-B cell stage D-J-gene segment rearrangement occurs within the IgH locus, followed by V- to D-J-gene segment combination (60). Prior to IgL

rearrangement, the heavy chain is expressed on the cell surface in complex with a surrogate light chain structure to confirm successful IgH rearrangement (61). During this time, and following IgL rearrangement, RAG1/2 expression is downregulated to prevent recombination of the immunoglobulin locus in the alternate allele (62, 63). In the event of an unproductive recombination, the cell can undergo a second recombination on the alternate IgH locus allele (60). The rearrangement of the light chain occurs in the pre-B cell (64). If unsuccessful recombination occurs at both alleles of the Ig $\kappa$  locus, then recombination will occur at the Ig $\lambda$  locus (65, 66). Unsuccessful recombination at both alleles of either the IgH or both IgL loci will result in cell death. The sequence of recombination events also exhibits some species specificity, for example, IgL development precedes IgH development in swine with no dependency on surrogate light chains in IgH selection (67).

Upon successful rearrangement of both IgH and IgL chains, immature B cells display fully formed immunoglobulin on their surface. Autoreactivity of the immunoglobulin is tested in the bone marrow and cells with autoreactive immunoglobulins will either undergo light chain editing or apoptosis, or become anergic (68-71). Immature B cells that are not self-reactive will express IgD in addition to IgM heavy chain constant regions through RNA splicing and traffic from the bone marrow to the spleen where maturation can occur (72, 73). Because it is possible for B cells with self-reactive immunoglobulins to escape central tolerance in the bone marrow, likely due to a lack of expression of tissue specific self-molecules, peripheral tolerance mechanisms also exist to clonally delete or induce anergy in autoantigen-specific B cells (74-76).

## **Further Immunoglobulin Diversification & Clonal Expansion**

Upon engagement of a mature B cell receptor with its antigen, the B cell can undergo a process known as somatic hypermutation, adding to the antibody diversity achieved through somatic recombination. This process is mediated by Activation-Induced Cytidine Deaminase (AID) (77, 78). First, inhibitory RNA is removed from AID through the action of an RNase (79). AID then binds to single-stranded DNA of at least twenty nucleotides that has been made accessible through transcription machinery, preferentially acting on cytidines that are surrounded by hotspot nucleotide motifs (79, 80). Hotspot motifs are particularly common in complementarity determining regions, introducing diversity in CDR1 and 2, while mutations occurring within framework regions are selected against due to potential disruption of antibody structure (81). The presence of the RNA base Uridine in DNA triggers either mismatch or base-excision DNA repair pathways. Mismatch repair is introduced when the proteins complex MSH2/6 recruits nucleases that resolve the mismatched uracil-guanidine pair, also introducing other nucleotides near the mismatch (82). Base excision occurs by the recognition of the uracil-guanidine pair by the uracil-DNA glycosylase, which will remove the uracil to create an abasic site that can be resolved during replication with the insertion of a random nucleotide (83). The introduction of mutations that result in higher antigen affinity of the antibody will allow for the selection of that B cell clone (84).

Finally, AID also mediates a second type of recombination that occurs in IgH known as class switch recombination (30, 77, 85, 86). This recombination event results in the expression of IgG, IgA or IgE depending on the type of antigenic stimulation (87). AID binds switch regions that are located in introns upstream of constant region genes (88, 89). One AID isoform binds to the switch region of the most upstream constant region gene ( $C\mu$  in non-class switched B cells), while another AID isoform binds to the switch region upstream of the constant region to be used,



introducing double-stranded nicks at both locations (88). Double-stranded break repair machinery then brings together the two switch regions, looping out and excising intervening sequences, resulting in an antibody of the same idiotype yet different effector functions (86). Together, somatic hypermutation and class switching mechanisms allow for the generation of an antibody repertoire with higher antigen affinity capable of specific effector functions.

## **High-throughput Sequencing and the Antibody Repertoire**

### **Early high throughput sequencing platforms**

High throughput sequencing (HTS) of messenger RNA and genomic DNA from B-cell populations can be used to assess the antibody repertoire (90). While studies using traditional Sanger sequencing were able to assess several hundred B cells, a greater sampling depth is required for a complete understanding of the antibody repertoire, which could theoretically contain over  $10^{13}$  unique sequences in humans due to the number of possible gene segment combinations and nucleotide additions (91, 92). Also referred to as massively parallel sequencing, HTS achieves greater sampling depth through the use of small nucleic acid templates (DNA or cDNA created from by rtPCR RNA) that are amplified and sequenced in parallel.

454 Pyrosequencing and Illumina were two of the first major HTS sequencing platforms (90). Both platforms require sequencing library preparation in which nucleic acid templates are fragmented and amplification of immunoglobulin specific primers can be performed. In 454 Pyrosequencing, templates are diluted and suspended in droplets containing all of the components necessary for template amplification. Within the droplet, amplified templates are affixed to beads and upon disruption of the droplet, beads are distributed in microwells, which are incubated cyclically with one base at a time. Pyrophosphates released as nucleotides are added to the template

strand by polymerase, ultimately generating light in reaction with the enzyme luciferase that is detected and recorded by sensors in the platform (90, 93).

The Illumina platform first requires the addition of an adapter to the 5' and 3' ends of the nucleic acid template. Varying preparation methods of platform-specific adapter sequences for sequencing library preparation have been shown to yield similar HTS results (94). Error and bias can be also assessed through the use of multiplex identifier adapters, synthetic immunoglobulin spike in sequences, and error correcting software (95). Sequences with adaptors are separated and amplified into clusters on the surface of a flow cell. Similar to Sanger sequencing, the Illumina platform utilizes fluorescently tagged terminated nucleotides. In Sanger sequencing, these labeled-terminator nucleotides are added in a mix to a DNA template, randomly terminating to ultimately generate a composite through electrophoresis and fluorescent imaging. In contrast, all four bases are added at once to the Illumina flow cell which is then imaged. Fluorescent labels and terminator molecules are then removed and a new cycle begins in an approach known as “sequencing by synthesis” (90, 93). 454 Pyrosequencing and Illumina sequencing platforms provide comparable data, but the Illumina platform provides greater sequencing depth than 454 Pyrosequencing (96). Due to its lower sequencing capacity and high cost per megabase pair, the 454 pyrosequencing platform has fallen out of favor and is no longer manufactured (97).

Initial high throughput sequencing studies focused on the IgH repertoire, however some studies have been done using both IgH and IgL, and pairing information was inferred through the association of gene segments by ranking sequence abundances of the independent chains (98). More recently, studies have attempted to more precisely pair IgH and IgL chains. In one approach, individual B cell emulsions are dispensed into microwells where they are lysed and mRNA of immunoglobulin chains is linked by microbead (99-101). cDNA synthesis and PCR amplification

are then performed on collected linked immunoglobulin sequences in preparation for Illumina sequencing, generating cDNA templates that span around 850 base pairs (99, 100, 102). In another approach, single cells are isolated onto PCR plates and forward and reverse primers for each well are tagged by columns and row (i.e. forward primers tagged uniquely for each column, reverse primers tagged uniquely for each row), identifying the mRNA sequences with the coordinates of the well (101, 103). An advantage of this approach is that heavy and light chains are not physically linked, resulting in template lengths that are more compatible with current HTS sequencing length capacities.

### **Considerations for Sequencing**

The B-cell targets for HTS are numerous. Large populations of B cells exist in several compartments such as bone marrow, spleen, lymph nodes and peripheral blood. In mice, any of these compartments can be easily sampled. B-cell repertoires are most frequently assessed in peripheral blood in humans, however bone marrow, cord blood and even formalin-fixed paraffin embedded lymph nodes have also been examined (104-106). B-cell populations and antibody idiotypes are tissue specific so sampling of multiple tissues provides a more complete view of the overall repertoire (107).

In addition to tissue heterogeneity, one must also consider B-cell subpopulations (108). An unsorted antibody repertoire may be skewed in favor of more transcriptionally active cells. For instance, there is a 500:5:2 ratio of mRNA levels from human plasma cells, memory B cells and Naïve B cells, respectively (109). Assessment of rearranged DNA provides information on gene segment sequences within the repertoire, while RNA provides information on the expression of immunoglobulin sequences within the repertoire (92).

Increased depth of sequencing can be accomplished using PCR amplification with primers specific to immunoglobulin gene segment family framework regions and constant regions. The drawback is that differences in primer hybridization efficiency may lead to primer bias of the repertoire (95). Primer selection was shown to have an impact on gene segment detection within the antibody repertoire (110). Rapid amplification of cDNA ends (RACE) is often used on the 5' end to amplify upstream of V-gene segments in combination with one or multiple constant region specific primers (92). This technique has been used as a gold standard in assessing primer bias in datasets that were created with gene-segment-specific primers (111). Gene-segment independent amplification high-throughput genome-wide translocation sequencing-adapted repertoire sequencing (HTGTS) recognizes signature recombination sites in DNA, obviating the need for degenerate V gene primers that may introduce bias (112).

Several correction methods can be implemented to reduce the impact that PCR and sequencing errors and biases can have on repertoire interpretation (98). Technical replicates can be used to filter sequencing reads that are not found in all replicates as a way to limit artificial diversity, however this comes at the cost of potentially excluding rare, genuine sequences and does not correct for highly reproducible sequencing errors (98, 113). Clustering of sequences based on similarity can be used to distinguish somatic variants and erroneous variants though the use of software such as MiXCR (114). The use of synthetic antibody spike-in standards can be used to quantify the degree of error and bias (95). The addition of unique molecular identifiers (UID) to nucleic acid templates during library preparation allows for more certainty in error correction by consensus sequence creation as sequences can be traced to a specific template (109, 115). Molecular amplification fingerprinting tags with both forward and reverse UIDs (95, 98) can also

be used. The ratio of forward and reverse tags can be used to normalize sequencing data with downstream bioinformatics software (95, 98).

### **Bioinformatic Assessment of Sequencing Data**

After sequencing, raw immunoglobulin reads are first filtered and quality screened using software such as pRESTO, IgBLAST, or CLC Genomics (116, 117) ([www.clcbio.com](http://www.clcbio.com)). Sequences are trimmed to remove error prone ends of sequencing reads and the reads that are assigned low Phred scores, or a high probability of an inaccurate base call, are removed (113). Quality cleaned reads can then be aligned to reference sequences to identify gene segment usage or further annotated with information such as CDR3 sequence composition and nucleotide mutation (113).

The international ImMunoGeneTics information system was established in 1989 as a immunogenetics reference which now contains numerous species-specific Ig and TCR gene segment databases and multiple repertoire annotation tools (118). The IMGT HighV-Quest tool was developed for the characterization of HTS antibody repertoire datasets through the submission of FASTA files of sequencing reads (119). Output files provide functional characterization of mapped sequencing reads in addition to gene segment usage, junctional analysis and mutation information. Due to the large amount of output information from HighV-Quest, many software packages have been developed to generate descriptive statistics and visualize data such as the excel based Immunoglobulin Analysis Tool (Ig-AT), Immunoglobulin (IGGalaxy; now expanded as Antigen Receptor Galaxy, ArGalaxy), Immune Explorer (IMEX) and analyzer-I (120-124).

Several comprehensive software packages have been developed independently of HighV-Quest. IgBLAST can be used for aligning immunoglobulin gene segments and characterizing

framework and complementarity determining regions (116). Repertoire specific toolkits such as Change-O have been developed for the identification of novel alleles, clonal characterization, mutation analysis and selection pressure through the use of Python command line script and custom R packages (125). MiXCR and iMonitor were designed as a toolkit for the analysis of Ig and TCR sequencing datasets in which reads are quality controlled, aligned to gene segments, assessed statistically, and visualized (114, 126). In addition to upstream quality control measures, both of these software packages utilize downstream structural analysis that corrects for PCR and sequencing errors. Data visualization includes gene segment usage and pairing, junctional insertion-deletion, SHM, and clonotype frequency, among other features. Similarly, ImmuneDiversity can process raw IgH sequencing data and provides output files for repertoire visualization (127).

Software has also been designed to assess specific details of immunoglobulins or the repertoire population such as affinity maturation (128, 129) and phylogenetic analysis such as IgTree, SONAR and Clonify (130-133). Tools such as IgDiscover and TIGER enable novel allele detection and genotype inferences in antibody repertoire datasets (134-136).

With the development of high-throughput single-cell analysis, software like sciReptor allow for the integration of common external software used for repertoire analysis with metadata such as patient characteristics and phenotypic characterizations from flow cytometry (137). In an effort to streamline the usage of multiple repertoire analysis tools, a uniform output format, VDJML, has been proposed (138). In a separate effort to improve efficiency of repertoire analysis, IgSimulator has been developed to allow researchers to generate simulated repertoires for testing of multiple bioinformatics pipelines to aid in experimental design (139). For example, use of a similar simulator program allowed for the development of a bioinformatics pipeline capable of

characterizing reads of only 70 bp in length (140). Additionally, biotechnology companies now offer antibody repertoire assessment of clinical isolates for research and diagnostics such as Adaptive biotechnologies ([www.adaptivebiotech.com](http://www.adaptivebiotech.com)) and iRepertoire, Inc. ([www.irepertoire.com](http://www.irepertoire.com)).

### **Murine Antibody Repertoire Studies**

While a large number of repertoire studies have already been performed using clinical isolates from patients and healthy volunteers, the ease of tissue collection from mice in vaccination and disease models can provide a broader landscape of antibody repertoire development and clues about the generation of antibody-mediated pathologies. Additionally, several mouse studies specifically tested sequencing methods and bioinformatic workflows (94, 95, 110, 112, 141).

Many studies assessed sorted B-cell populations such as pre/pro-B cells, long lived plasma cells, follicular B cells, marginal B cells, B1a B cells, and B-2 B cells (142-146). While studies have isolated B cells from the lymph node (147) or peritoneum (145), bone marrow and the spleen are the most commonly sampled tissues. These studies have shown that the antibody repertoire is B-cell population, and tissue-specific. Gene segment usage is also strain specific (148). Despite inbreeding, a high level of variation in CDR3 sequences also occurs among individual mice (110). A comparison of laboratory and wild mice revealed higher levels of serum immunoglobulins, lower absolute counts of splenic B cells and reduced cytokine responses in wild mice, likely driven by higher disease burden (149).

The antibody repertoire in mice has been assessed in response to specific antigens (95, 143, 150-152) and also in germ-free animals (145). Through challenging mice from different genetic backgrounds with multiple antigens and assessing multiple B cell populations, Grieff et al. found

that the antibody repertoire is largely predetermined by genetic background and antigen exposure, although stochastic variation was still seen (152). Common antigen exposure between individual animals results in a higher number of shared, or “public, clones” between all animals in the treatment group (150). These data lead to the understanding of repertoire dynamics and a higher resolution predictions of which B cells will respond to antigen (150).

### **Antibody Repertoire Studies Enrich Understanding of Basic Human B-cell Biology**

Developmentally, the repertoire of human cord blood shows preferential gene segment usage in Heavy, Kappa and Lambda chains (104). Class switching and somatic hypermutation increased with age in early childhood (153). The bone marrow plasma cell repertoire in adolescent patients does not change much between two collection points 6.5 years apart (154). However, antibody repertoires do change in response to vaccination in both young (aged 19-45) and old (aged 70-89) volunteers (155). Aged volunteers had fewer mutations, longer CDR3, and differ in immunoglobulin class distribution (155). Other studies have reported that the repertoire varies among individuals but is similar in young and old patients (108, 156). Therefore, it appears that repertoires vary considerably among individuals in the population.

The antibody repertoire shows tissue specificity. Mucosal tissues (lung, small intestine, stomach) had more mutations and longer CDR3 than other lymphoid tissues (Lymph node, Tonsil, Spleen, Thymus) (157). Differences in gene segment usage across B cell subsets within tissues also occur (157). Assessment of large Pre-B, immature, transitional and naïve B cells from bone marrow and the peripheral blood of healthy adult donors revealed that the transitional B-cell repertoire differs from B cells at other developmental stages, indicating that transitional B cells may not merely be precursors to naïve B cells (108).



Although the number of B cell idiotypes seems theoretically unlimited, a number of studies on B-cell Ig segment use have shown that the number of idiotypes may be more limited. For example, a strong bias in IgH D-J pairing, independent of V-gene segment pairing, has been reported (158). Beneichou et al. showed preferential reading frame usage in human D-gene segments also resulting in less diversity than theoretically possible (159). In another study, reading frame bias was not observed among D-gene segments, although preferential gene segment usage was noted (160). In violation of the 12/23 rule, rare non-canonical V(DD)J rearrangements have been detected within human peripheral blood, which would promote more idiotypic diversity (161).

Although CDR3 usage is varied between individuals, V(D)J-gene segment combinations show overlap (162, 163). Inter-individual sampling of B cells isolated from peripheral blood of 10 healthy individuals, showed strong Pearson correlations of most abundant IgH V-J combinations; ranging between 0.63 and 0.96 (163). Repeated sampling of one participant once weekly for five weeks found strong consistency in the individual's V-J combinations (Pearson correlations; 0.983-0.999) (163). Arnaout et al. found that individual gene segment usage (V-, D-, and J-genes) was highly correlated between individuals ( $R^2 = 0.91, 0.90$  and  $0.97$ , V, D, J, respectively) (158). Both Ig $\kappa$  and Ig $\lambda$  have a high number of publically shared clones as well (164, 165). Since many of these studies assessed the repertoire of only a few individuals, more sampling will be necessary to understand the Ig gene repertoire of the population since there is such a large diversity in Ig gene segment usage and because of the number of allelic variants within immunoglobulin loci (166, 167).

## **Antibody Repertoire in Human Disease**

Antibody repertoire studies have been quickly adopted to study human vaccine responses and autoimmune diseases (168-172), B-cell malignancies (173-177), transplantation responses (178), and allergic responses (179). For example, studies on the impact of vaccination on the antibody repertoire have been conducted for influenza (173, 180-183), hepatitis (181, 182), tetanus (167, 182, 184, 185), meningococcus (182, 186) and the anthrax vaccine (187). Of note, assessment of hepatitis B vaccination revealed cross-reactive antibodies from B cells with low affinity to the vaccine in the first of three vaccine administrations, with higher affinity clones successfully isolated in subsequent vaccinations (188). Increased class switching and more abundant clones were detected in the peripheral blood of patients receiving intradermally administered trivalent inactivated seasonal influenza vaccine compared to patients receiving nasally administered live attenuated vaccine (189).

A framework has been proposed for “immune-repertoire based finger printing” for early detection of disease (190). Immune repertoire analysis could have diagnostic potential for the detection of minimally residual disease in B-cell malignancies (174). Antibody repertoire assessment may also have prognostic utility as low levels of antibody repertoire diversity have been correlated with mortality in patients infected with avian influenza (191). Moreover, “reverse vaccinology” HTS of antibody repertoires can be used to design protective antibodies (192).

In conclusion, the rapid application of repertoire sequencing directly in the clinic setting, increased understanding of basic repertoire dynamics, optimization of sequencing technologies, and the development of bioinformatics pipelines to identify signals that immunoglobulin repertoire analysis as a rich area of research.

## **Spaceflight and the Adaptive Immune Response**

### **Stressors in Spaceflight and Terrestrial Analogs of Spaceflight**

Spaceflight presents a unique set of physiological challenges because it impacts the musculoskeletal, cardiovascular, sensory-motor, and immune systems of astronauts (193). Suppression of adaptive immune responses has been noted such as altered T- and B-lymphocyte subpopulation distribution (194-198), with reduced proliferation (195-198) and altered cytokine profiles against common immunogens (195, 196, 199, 200).

To characterize the effects of spaceflight, several *in vitro* and physiological stress analogs have been employed. Variables that influence immune homeostasis include microgravity and radiation, which are specific to the spaceflight environment, in addition to mission-specific variables that include the acute stress of launch and landing, and chronic variables such as psychological stress and altered sleep, diet, and exercise (199). Antiorthostatic suspension (AOS) is an *in vivo*, ground-based mouse model in which weight on the hind limbs is unloaded through suspension and animals are kept with a head-down tilt. AOS induces changes such as bone and muscle loss in the hind limbs, fluid shift toward the head, and immune cell changes (199, 201, 202). In addition to physical stress, this model also induces physiological stress that stimulates the release of stress hormones that disrupt immune homeostasis (199, 201, 202).

In some studies, unloading has been paired with launch and landing stress through centrifugations (199, 203). Human bedrest (hypokinesia) studies that employ a head-down tilt have provided mixed findings and have been shown to more closely resemble neuroendocrine and immune responses of short-duration astronauts when paired with simulated launch and landing stress (199, 204). Short term gravitational vector changes from parabolic flights have been used in both animal and tissue culture models (205, 206). Simulated microgravity in tissue-culture systems

has also been induced by clinostats and rotating bioreactors <http://www.adaptivebiotech.com/immunoseq/products/service> in which the gravity vector constantly changes (199, 207).

Spaceflight outside of low-Earth orbit will increase exposure several types of radiation, such as gamma-rays, X-rays, protons and heavy ions, that have been tested in models either independently or in addition to simulated microgravity (208-210). In both spaceflight studies and ground-based analogs, immune cell populations are often assessed phenotypically by flow cytometric analysis of cell surface markers, and functionally through measurement of proliferation and cytokine profiles in response to stimulation (211, 212).

### **The Adaptive Immune Response in Space**

T and B lymphocytes are adaptive immune cells that express antigen-specific receptors. Although indistinguishable through microscopy, they can be distinguished by the presence of cell surface cluster of differentiation (CD) molecules, which can be labeled by CD-specific antibodies and assessed by flow cytometry or immunohistochemistry imaging (213). Three main types of T cells are cytotoxic T cells, ( $CD8^+$ ), helper T cells ( $CD4^+$ ), and regulatory T cells,  $CD4^+$  T cells that express the transcription factor FOXP3 (213). B cells are broadly identified through CD19 expression (213). While distinct T- and B-cell subpopulations can be further characterized through additional markers, such distinctions are not often made in assessing the adaptive immune response in space.

The physiological changes associated with space flight are manifested in many ways in the immune system. For example, lymphocyte populations change in response to spaceflight (214-221) or as a result of AOS (211, 222). Decreases in lymphocytes compared to controls also can

occur (211, 212, 214-216, 222-226), although other studies contradict those data and have shown no significant changes or organ-specific changes (227, 228). The T-cell subpopulation ratio of CD4<sup>+</sup> helper to CD8<sup>+</sup> cytotoxic T cells has been shown to decrease, which could also result in reduced B-cell activation (219).

The mass of lymphoid organs is often reduced by space flight or AOS (203, 211, 212, 216, 223-227). Atrophy observed in the lymphoid follicles of the spleen (226) suggests lower lymphocyte numbers. Increased DNA double strand breaks were detected more frequently in the spleen and thymus of hindlimb unloaded animals, suggestive of increased apoptosis to explain why cell numbers would decrease (222). Cortical thymocytes are sensitive to stress glucocorticoids (stress hormones), which are elevated by spaceflight and AOS (200, 225, 229, 230). Additionally, upregulation of 5-lipoxygenase was observed in human lymphocyte cultures aboard the International Space Station and was correlated with higher levels of DNA fragmentation (231). Spaceflight altered the phenotype of immune cells in the bone marrow differentiated in space. This change in immune cell phenotypes, suggests a direct effect of space flight on bone marrow cells (232). A reduced number of common lymphoid progenitors, from which T and B cells arise, was found in the bone marrow of hindlimb unloaded mice compared to controls (233).

Functionally, T-cell populations in spaceflight conditions have been shown to have lower expression of early activation genes and cell surface markers (205, 217, 234-237), reduced proliferative activity upon mitogen challenge (212, 216, 237-243) and altered cytokine expression profiles (212, 217, 219, 240, 243-245). Interleukin (IL) 6 is an inflammatory cytokine produced by macrophages that promotes lymphocyte activation and antibody production. IL-6 was found to be increased in post-landing splenocytes of space flown mice in response to LPS, however other rodent studies examining splenocyte production of IL-6 post-landing spaceflight (235, 245) or

immediately after AOS (211) showed no increase in IL-6. In astronauts, IL-6 spikes at both launch and landing. The response paralleled spikes in cortisol levels (200) and suggests that launch and landing stresses are having an impact. As with rodents, the changes in IL-6 were not consistent in astronauts as IL-6 was below the detection threshold regardless of when it was measured during other space flights (246). IL-6 expression was reduced in activated human lymphocyte tissue cultures compared to ground controls (247).

Another cytokine assessed by multiple studies is tumor necrosis factor (TNF)  $\alpha$ , an inflammatory cytokine produced by macrophages and T cells. TNF- $\alpha$  was slightly elevated above the detection threshold in all in-flight astronaut plasma samples, whereas all pre- and post-flight samples had TNF- $\alpha$  levels that were below the detection threshold (246). In contrast, TNF- $\alpha$  secretion and transcription decreased in activated rodent splenocytes (223, 235, 240) and no significant differences were seen in activated rodent splenocytes from mice treated with and without AOS (211).

Changes in other cytokines have also been observed. Decreases in T-cell IL-2 (196, 212, 236, 240), IL-5 (245), and IFN- $\gamma$  (236, 243, 244) occur as a result of spaceflight. The changes in cytokine expression profiles may be age dependent, as blastogenesis and IFN- $\gamma$  production by rat splenocytes was observed in dams but not pups flown in space (243).

Other functional changes, such as reduced locomotion of lymphocytes through collagen in response to phytohemagglutinin stimulation (248, 249) and reduced signal transduction, have been observed in human lymphocytes grown in rotational bioreactors (250). CD4<sup>+</sup> T cells activated aboard the International Space Station had reduced expression of miR-21, a micro RNA that had predicted targets corresponding to 16 of the 85 genes differentially expressed in spaceflight compared to ground controls (251).

## **Current understanding of spaceflight and the antibody repertoire**

Some aspects of the impact of spaceflight on lymphocyte receptor repertoires have been explored. Activated human lymphocyte tissue cultures showed reduced IgM production aboard the International Space Station and in rotating bioreactors compared to static controls (247). Additionally, hypergravity modeled by centrifugation in newborn C57BL/6 mice was shown to modify the repertoire of T cell receptors, which are assembled similarly to immunoglobulins (252). The splenic IgH chains of the amphibian *Pleurodeles waltl* were examined 10 days after being flown aboard Mir for 5 months (253). VH families and IgM and IgY were quantified, revealing an increased level of IgY in flight animals compared to ground controls and modified IgM VH gene-family usage (253, 254). In a study of the impact of spaceflight on the immunoglobulin repertoire development during ontogeny of *P. waltl*, spaceflight resulted in higher levels of IgM transcripts (255). Changes to recombination machinery occurred in animals subjected to hypergravity including increased mRNA levels of RAG1, decreased G base insertion, and increased T base insertion (252). There was little overlap between gene segment combinations of control animals and animals subjected to centrifugation (252).

In summary, the physiological changes associated with space flight are quite comprehensive. Changes in differentiation, population distribution, trafficking, as well as functional cytokine responses all can affect B-cell function directly or indirectly. Given these varied mechanisms one might expect some impact on B-cell differentiation; an important part of which is the recombination of immunoglobulin genes needed for the synthesis of the antibody receptors that define B cells.

## **Objective**

This work tested the hypothesis that space flight affects the immunoglobulin repertoire of unimmunized C57BL/6 mice. This hypothesis was tested through the characterization of the Igκ immunoglobulin repertoire from high-throughput sequencing datasets. First, sample preparation methods and bioinformatics workflows were validated (Chapter 2). Next, the immunoglobulin repertoire was more fully characterized by assessing individual segment usage, gene segment combinations and junctions among pooled biological replicates (Chapter 3). Finally, the Igκ immunoglobulin repertoires of mice flown aboard the International Space Station as a part of the Rodent Research One validation flight were compared to the immunoglobulin repertoires of ground control animals (Chapter 4). This work will be continued in the future to study antibody repertoire dynamics in response to vaccination within the setting of a spaceflight analog.



## References

1. von Behring, E., and S. Kitasato. 1890. The mechanism of diphtheria immunity and tetanus immunity in animals. *Dtsch. Med. Wochenschr.* 16: 1113-1114.
2. von Behring, E. 1890. Studies on the mechanism of immunity to diphtheria in animals. *Dtsch. Med. Wochenschr.* 16: 1145-1148.
3. Ehrlich, P. 1900. Croonian: On immunity with special reference to cell life. *Proc. R. Soc. Lond.* 66: 424-448.
4. Fagraeus, A. 1947. Plasma cellular reaction and its relation to the formation of antibodies in vitro. *Nature* 159: 499.
5. Nossal, G. J., and J. Lederberg. 1958. Antibody production by single cells. *Nature* 181: 1419-1420.
6. Harris, T. N., E. Grimm, E. Mertens, and W. E. Ehrlich. 1945. The role of the lymphocyte in antibody formation. *J. Exp. Med.* 81: 73-83.
7. Cooper, M. D., R. D. Peterson, and R. A. Good. 1965. Delineation of th thymic and bursal lymphoid systems in the chicken. *Nature* 205: 143-146.
8. Porter, R. R. 1959. The hydrolysis of rabbit  $\gamma$ -globulin and antibodies with crystalline papain. *Biochem J.* 73: 119-126.
9. Nisonoff, A., F. C. Wissler, and L. N. Lipman. 1960. Properties of the major component of a peptic digest of rabbit antibody. *Science* 132: 1770-1771.
10. Lesk, A. M., and C. Chothia. 1988. Elbow motion in the immunoglobulins involves a molecular ball-and-socket joint. *Nature* 335: 188-190.

11. Edelman, G. M., B. Benacerraf, Z. Ovary, and M. D. Poulik. 1961. Structural differences among antibodies of different specificities. *Proceedings of the National Academy of Sciences of the United States of America* 47: 1751-1758.
12. Hilschmann, N., and L. C. Craig. 1965. Amino acid sequence studies with Bence-Jones proteins. *Proc. Natl. Acad. Sci.* 53: 1403-1409.
13. Titani, K., E. Whitley, Jr., L. Avogardo, and F. W. Putnam. 1965. Immunoglobulin structure: partial amino acid sequence of a Bence Jones protein. *Science* 149: 1090-1092.
14. Milstein, C. 1967. Variations in the C-terminal half of immunoglobulin gamma-chains. *Biochem J.* 104: 28-30.
15. Hill, R. L., R. Delaney, R. E. Fellows, and H. E. Lebovitz. 1966. The evolutionary origins of the immunoglobulins. *Proc. Natl. Acad. Sci.* 56: 1762-1769.
16. Black, C. A. 1997. A brief history of the discovery of the immunoglobulins and the origin of the modern immunoglobulin nomenclature. *Immunol. Cell Biol.* 75: 65-68.
17. Swan, D., P. D'Eustachio, L. Leinwand, J. Seidman, D. Keithley, and F. H. Ruddle. 1979. Chromosomal assignment of the mouse kappa light chain genes. *Proc. Natl. Acad. Sci.* 76: 2735-2739.
18. D'Eustachio, P., D. Pravtcheva, K. Marcu, and F. H. Ruddle. 1980. Chromosomal location of the structural gene cluster encoding murine immunoglobulin heavy chains. *J. Exp. Med.* 151: 1545-1550.
19. D'Eustachio, P., A. L. Bothwell, T. K. Takaro, D. Baltimore, and F. H. Ruddle. 1981. Chromosomal location of structural genes encoding murine immunoglobulin lambda light chains. Genetics of murine lambda light chains. *J. Exp. Med.* 153: 793-800.

20. Dreyer, W. J., and J. C. Bennett. 1965. The molecular basis of antibody formation: a paradox. *Proc. Natl. Acad. Sci.* 54: 864-869.
21. Brack, C., and S. Tonegawa. 1977. Variable and constant parts of the immunoglobulin light chain gene of a mouse myeloma cell are 1250 nontranslated bases apart. *Proc. Natl. Acad. Sci.* 74: 5652-5656.
22. Seidman, J. G., and P. Leder. 1978. The arrangement and rearrangement of antibody genes. *Nature* 276: 790-795.
23. Hozumi, N., and S. Tonegawa. 1976. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl. Acad. Sci.* 73: 3628-3632.
24. Tonegawa, S., A. M. Maxam, R. Tizard, O. Bernard, and W. Gilbert. 1978. Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain. *Proc. Natl. Acad. Sci.* 75: 1485-1489.
25. Bernard, O., N. Hozumi, and S. Tonegawa. 1978. Sequences of mouse immunoglobulin light chain genes before and after somatic changes. *Cell* 15: 1133-1144.
26. Sakano, H., K. Huppi, G. Heinrich, and S. Tonegawa. 1979. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* 280: 288-294.
27. Maki, R., A. Traunecker, H. Sakano, W. Roeder, and S. Tonegawa. 1980. Exon shuffling generates an immunoglobulin heavy chain gene. *Proc. Natl. Acad. Sci.* 77: 2138-2142.
28. Tonegawa, S. 1976. Reiteration frequency of immunoglobulin light chain genes: further evidence for somatic generation of antibody diversity. *Proc. Natl. Acad. Sci.* 73: 203-207.
29. Brack, C., M. Hirama, R. Lenhard-Schuller, and S. Tonegawa. 1978. A complete immunoglobulin gene is created by somatic recombination. *Cell* 15: 1-14.

30. Sakano, H., R. Maki, Y. Kurosawa, W. Roeder, and S. Tonegawa. 1980. Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. *Nature* 286: 676-683.
31. Early, P., H. Huang, M. Davis, K. Calame, and L. Hood. 1980. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. *Cell* 19: 981-992.
32. Wu, T. T., and E. A. Kabat. 1970. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *The Journal of experimental medicine* 132: 211-250.
33. Singer, S. J., and R. F. Doolittle. 1966. Antibody active sites and immunoglobulin molecules. *Science* 153: 13-25.
34. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575-581.
35. Lewis, S., A. Gifford, and D. Baltimore. 1985. DNA elements are asymmetrically joined during the site-specific recombination of kappa immunoglobulin genes. *Science* 228: 677-685.
36. Ramsden, D. A., J. F. McBlane, D. C. van Gent, and M. Gellert. 1996. Distinct DNA sequence and structure requirements for the two steps of V(D)J recombination signal cleavage. *EMBO J.* 15: 3197-3206.
37. van Gent, D. C., D. A. Ramsden, and M. Gellert. 1996. The RAG1 and RAG2 proteins establish the 12/23 rule in V(D)J recombination. *Cell* 85: 107-113.
38. Schatz, D. G., M. A. Oettinger, and D. Baltimore. 1989. The V(D)J recombination activating gene, RAG-1. *Cell* 59: 1035-1048.

39. Oettinger, M. A., D. G. Schatz, C. Gorka, and D. Baltimore. 1990. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science* 248: 1517-1523.
40. Sawchuk, D. J., F. Weis-Garcia, S. Malik, E. Besmer, M. Bustin, M. C. Nussenzweig, and P. Cortes. 1997. V(D)J recombination: modulation of RAG1 and RAG2 cleavage activity on 12/23 substrates by whole cell extract and DNA-bending proteins. *J. Exp. Med.* 185: 2025-2032.
41. van Gent, D. C., K. Hiom, T. T. Paull, and M. Gellert. 1997. Stimulation of V(D)J cleavage by high mobility group proteins. *EMBO J.* 16: 2665-2670.
42. Schlissel, M., A. Constantinescu, T. Morrow, M. Baxter, and A. Peng. 1993. Double-strand signal sequence breaks in V(D)J recombination are blunt, 5'-phosphorylated, RAG-dependent, and cell cycle regulated. *Genes Dev.* 7: 2520-2532.
43. Agrawal, A., and D. G. Schatz. 1997. RAG1 and RAG2 form a stable postcleavage synaptic complex with DNA containing signal ends in V(D)J recombination. *Cell* 89: 43-53.
44. Ramsden, D. A., and M. Gellert. 1998. Ku protein stimulates DNA end joining by mammalian DNA ligases: a direct role for Ku in repair of DNA double-strand breaks. *EMBO J.* 17: 609-614.
45. Li, Z., T. Otevrel, Y. Gao, H. L. Cheng, B. Seed, T. D. Stamato, G. E. Taccioli, and F. W. Alt. 1995. The XRCC4 gene encodes a novel protein involved in DNA double-strand break repair and V(D)J recombination. *Cell* 83: 1079-1089.
46. Grawunder, U., D. Zimmer, P. Kulesza, and M. R. Lieber. 1998. Requirement for an interaction of XRCC4 with DNA ligase IV for wild-type V(D)J recombination and DNA double-strand break repair in vivo. *J. Biol. Chem.* 273: 24708-24714.

47. Moshous, D., I. Callebaut, R. de Chasseval, B. Corneo, M. Cavazzana-Calvo, F. Le Deist, I. Tezcan, O. Sanal, Y. Bertrand, N. Philippe, A. Fischer, and J. P. de Villartay. 2001. Artemis, a novel DNA double-strand break repair/V(D)J recombination protein, is mutated in human severe combined immune deficiency. *Cell* 105: 177-186.
48. Lewis, S. M. 1994. P nucleotide insertions and the resolution of hairpin DNA structures in mammalian cells. *Proc. Natl. Acad. Sci.* 91: 1332-1336.
49. Landau, N. R., D. G. Schatz, M. Rosa, and D. Baltimore. 1987. Increased frequency of N-region insertion in a murine pre-B-cell line infected with a terminal deoxynucleotidyl transferase retroviral expression vector. *Mol. Cell. Biol.* 7: 3237-3243.
50. Lieber, M. R., J. E. Hesse, K. Mizuuchi, and M. Gellert. 1988. Lymphoid V(D)J recombination: nucleotide insertion at signal joints as well as coding joints. *Proc. Natl. Acad. Sci.* 85: 8588-8592.
51. Kirchgessner, C. U., C. K. Patil, J. W. Evans, C. A. Cuomo, L. M. Fried, T. Carter, M. A. Oettinger, and J. M. Brown. 1995. DNA-dependent kinase (p350) as a candidate gene for the murine SCID defect. *Science* 267: 1178-1183.
52. Peterson, S. R., A. Kurimasa, M. Oshimura, W. S. Dynan, E. M. Bradbury, and D. J. Chen. 1995. Loss of the catalytic subunit of the DNA-dependent protein kinase in DNA double-strand-break-repair mutant mammalian cells. *Proc. Natl. Acad. Sci.* 92: 3171-3174.
53. Pergola, F., M. Z. Zdzienicka, and M. R. Lieber. 1993. V(D)J recombination in mammalian cell mutants defective in DNA double-strand break repair. *Mol. Cell. Biol.* 13: 3464-3471.
54. Taccioli, G. E., G. Rathbun, E. Oltz, T. Stamato, P. A. Jeggo, and F. W. Alt. 1993. Impairment of V(D)J recombination in double-strand break repair mutants. *Science* 260: 207-210.

55. Gu, Y., S. Jin, Y. Gao, D. T. Weaver, and F. W. Alt. 1997. Ku70-deficient embryonic stem cells have increased ionizing radiosensitivity, defective DNA end-binding activity, and inability to support V(D)J recombination. *Proc. Natl. Acad. Sci.* 94: 8076-8081.
56. Li, Y. S., K. Hayakawa, and R. R. Hardy. 1993. The regulated expression of B lineage associated genes during B cell differentiation in bone marrow and fetal liver. *J. Exp. Med.* 178: 951-960.
57. Allman, D., J. Li, and R. R. Hardy. 1999. Commitment to the B lymphoid lineage occurs before DH-JH recombination. *J. Exp. Med.* 189: 735-740.
58. Bangs, L. A., I. E. Sanz, and J. M. Teale. 1991. Comparison of D, JH, and junctional diversity in the fetal, adult, and aged B cell repertoires. *J. Immunol.* 146: 1996-2004.
59. Feeney, A. J. 1990. Lack of N regions in fetal and neonatal mouse immunoglobulin V-D-J junctional sequences. *J. Exp. Med.* 172: 1377-1390.
60. Alt, F. W., G. D. Yancopoulos, T. K. Blackwell, C. Wood, E. Thomas, M. Boss, R. Coffman, N. Rosenberg, S. Tonegawa, and D. Baltimore. 1984. Ordered rearrangement of immunoglobulin heavy chain variable region segments. *EMBO J.* 3: 1209-1219.
61. Nishimoto, N., H. Kubagawa, T. Ohno, G. L. Gartland, A. K. Stankovic, and M. D. Cooper. 1991. Normal pre-B cells express a receptor complex of mu heavy chains and surrogate light-chain proteins. *Proc. Natl. Acad. Sci.* 88: 6284-6288.
62. Loffert, D., A. Ehlich, W. Muller, and K. Rajewsky. 1996. Surrogate light chain expression is required to establish immunoglobulin heavy chain allelic exclusion during early B cell development. *Immunity* 4: 133-144.

63. Grawunder, U., T. M. Leu, D. G. Schatz, A. Werner, A. G. Rolink, F. Melchers, and T. H. Winkler. 1995. Down-regulation of RAG1 and RAG2 gene expression in preB cells after functional immunoglobulin heavy chain rearrangement. *Immunity* 3: 601-608.
64. Reth, M., E. Petrac, P. Wiese, L. Lobel, and F. W. Alt. 1987. Activation of V kappa gene rearrangement in pre-B cells follows the expression of membrane-bound immunoglobulin heavy chains. *EMBO J.* 6: 3299-3305.
65. Hieter, P. A., S. J. Korsmeyer, T. A. Waldmann, and P. Leder. 1981. Human immunoglobulin kappa light-chain genes are deleted or rearranged in lambda-producing B cells. *Nature* 290: 368-372.
66. Coleclough, C., R. P. Perry, K. Karjalainen, and M. Weigert. 1981. Aberrant rearrangements contribute significantly to the allelic exclusion of immunoglobulin gene expression. *Nature* 290: 372-378.
67. Sinkora, M., J. Sinkorova, and K. Stepanova. 2017. Ig Light Chain Precedes Heavy Chain Gene Rearrangement during Development of B Cells in Swine. *J. Immunol.* 198: 1543-1552.
68. Chen, C., Z. Nagy, M. Z. Radic, R. R. Hardy, D. Huszar, S. A. Camper, and M. Weigert. 1995. The site and stage of anti-DNA B-cell deletion. *Nature* 373: 252-255.
69. Melamed, D., R. J. Benschop, J. C. Cambier, and D. Nemazee. 1998. Developmental regulation of B lymphocyte immune tolerance compartmentalizes clonal selection from receptor selection. *Cell* 92: 173-182.
70. Gay, D., T. Saunders, S. Camper, and M. Weigert. 1993. Receptor editing: an approach by autoreactive B cells to escape tolerance. *J. Exp. Med.* 177: 999-1008.



71. Pike, B. L., A. W. Boyd, and G. J. Nossal. 1982. Clonal anergy: the universally anergic B lymphocyte. *Proc. Natl. Acad. Sci.* 79: 2013-2017.
72. Yuan, D., and P. L. Witte. 1988. Transcriptional regulation of mu and delta gene expression in bone marrow pre-B and B lymphocytes. *J. Immunol.* 140: 2808-2814.
73. Loder, F., B. Mutschler, R. J. Ray, C. J. Paige, P. Sideras, R. Torres, M. C. Lamers, and R. Carsetti. 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: 75-89.
74. Sidman, C. L., and E. R. Unanue. 1975. Receptor-mediated inactivation of early B lymphocytes. *Nature* 257: 149-151.
75. Goodnow, C. C., J. Crosbie, S. Adelstein, T. B. Lavoie, S. J. Smith-Gill, R. A. Brink, H. Pritchard-Briscoe, J. S. Wotherspoon, R. H. Loblay, K. Raphael, and et al. 1988. Altered immunoglobulin expression and functional silencing of self-reactive B lymphocytes in transgenic mice. *Nature* 334: 676-682.
76. Russell, D. M., Z. Dembic, G. Morahan, J. F. Miller, K. Burki, and D. Nemazee. 1991. Peripheral deletion of self-reactive B cells. *Nature* 354: 308-311.
77. Muramatsu, M., K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, and T. Honjo. 2000. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 102: 553-563.
78. Petersen-Mahrt, S. K., R. S. Harris, and M. S. Neuberger. 2002. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* 418: 99-103.

79. Bransteitter, R., P. Pham, M. D. Scharff, and M. F. Goodman. 2003. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. *Proc. Natl. Acad. Sci.* 100: 4102-4107.
80. Pham, P., R. Bransteitter, J. Petruska, and M. F. Goodman. 2003. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* 424: 103-107.
81. Foster, S. J., T. Dorner, and P. E. Lipsky. 1999. Somatic hypermutation of V $\kappa$ J $\kappa$  rearrangements: targeting of RGYW motifs on both DNA strands and preferential selection of mutated codons within RGYW motifs. *Eur. J. Immunol.* 29: 4011-4021.
82. Roa, S., Z. Li, J. U. Peled, C. Zhao, W. Edelmann, and M. D. Scharff. 2010. MSH2/MSH6 complex promotes error-free repair of AID-induced dU:G mispairs as well as error-prone hypermutation of A:T sites. *PloS one* 5: e11182.
83. Di Noia, J., and M. S. Neuberger. 2002. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature* 419: 43-48.
84. Davie, J. M., and W. E. Paul. 1973. Immunological maturation. Preferential proliferation of high-affinity precursor cells. *J. Exp. Med.* 137: 201-204.
85. Revy, P., T. Muto, Y. Levy, F. Geissmann, A. Plebani, O. Sanal, N. Catalan, M. Forveille, R. Dufourcq-Labelouse, A. Gennery, I. Tezcan, F. Ersoy, H. Kayserili, A. G. Ugazio, N. Brousse, M. Muramatsu, L. D. Notarangelo, K. Kinoshita, T. Honjo, A. Fischer, and A. Durandy. 2000. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell* 102: 565-575.
86. Chaudhuri, J., and F. W. Alt. 2004. Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. *Nat. Rev. Immunol.* 4: 541-552.

87. Ehrhardt, R. O., G. R. Harriman, J. K. Inman, N. Lycke, B. Gray, and W. Strober. 1996. Differential activation requirements of isotype-switched B cells. *Eur. J Immunol.* 26: 1926-1934.
88. Jung, S., K. Rajewsky, and A. Radbruch. 1993. Shutdown of class switch recombination by deletion of a switch region control element. *Science* 259: 984-987.
89. Yu, K., F. T. Huang, and M. R. Lieber. 2004. DNA substrate length and surrounding sequence affect the activation-induced deaminase activity at cytidine. *J. Biol. Chem.* 279: 6496-6500.
90. Benichou, J., R. Ben-Hamo, Y. Louzoun, and S. Efroni. 2012. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135: 183-191.
91. Arnaout, R. A. 2005. Specificity and overlap in gene segment-defined antibody repertoires. *BMC Genomics* 6: 148.
92. Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32: 158-168.
93. Boyd, S. D., and S. A. Joshi. 2014. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. *Microbiol. Spectr.* 2: 1-13.
94. Menzel, U., V. Greiff, T. A. Khan, U. Haessler, I. Hellmann, S. Friedensohn, S. C. Cook, M. Pogson, and S. T. Reddy. 2014. Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PloS one* 9: e96727.

95. Khan, T. A., S. Friedensohn, A. R. Gorter de Vries, J. Straszewski, H. J. Ruscheweyh, and S. T. Reddy. 2016. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* 2: e1501371.
96. Bashford-Rogers, R. J., A. L. Palser, S. F. Idris, L. Carter, M. Epstein, R. E. Callard, D. C. Douek, G. S. Vassiliou, G. A. Follows, M. Hubank, and P. Kellam. 2014. Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunol.* 15: 29.
97. Wardemann, H. B., C.E. 2017. Novel Approaches to Analyze Immunoglobulin Repertoires. *Trends Immunol.* 38: 471-482.
98. Friedensohn, S., T. A. Khan, and S. T. Reddy. 2017. Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires. *Trends Biotechnol.* 35: 203-214.
99. DeKosky, B. J., G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke, T. Dorner, S. F. Andrews, P. C. Wilson, S. P. Hunicke-Smith, C. G. Willson, A. D. Ellington, and G. Georgiou. 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 31: 166-169.
100. DeKosky, B. J., T. Kojima, A. Rodin, W. Charab, G. C. Ippolito, A. D. Ellington, and G. Georgiou. 2015. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* 21: 86-91.
101. Busse, C. E., I. Czogiel, P. Braun, P. F. Arndt, and H. Wardemann. 2014. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J. Immunol.* 44: 597-603.

102. McDaniel, J. R., B. J. DeKosky, H. Tanno, A. D. Ellington, and G. Georgiou. 2016. Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat. Protoc.* 11: 429-442.
103. Murugan, R., K. Imkeller, C. E. Busse, and H. Wardemann. 2015. Direct high-throughput amplification and sequencing of immunoglobulin genes from single human B cells. *Eur. J. Immunol* 45: 2698-2700.
104. Prabakaran, P., W. Chen, M. G. Singarayan, C. C. Stewart, E. Streaker, Y. Feng, and D. S. Dimitrov. 2012. Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* 64: 337-350.
105. Tabibian-Keissar, H., G. Schibby, M. Michaeli, A. Rakovsky-Shapira, N. Azogui-Rosenthal, D. K. Dunn-Walters, K. Rosenblatt, R. Mehr, and I. Barshack. 2013. PCR amplification and high throughput sequencing of immunoglobulin heavy chain genes from formalin-fixed paraffin-embedded human biopsies. *Exp. Mol. Pathol.* 94: 182-187.
106. Tabibian-Keissar, H., L. Hazanov, G. Schiby, N. Rosenthal, A. Rakovsky, M. Michaeli, G. L. Shahaf, Y. Pickman, K. Rosenblatt, D. Melamed, D. Dunn-Walters, R. Mehr, and I. Barshack. 2016. Aging affects B-cell antigen receptor repertoire diversity in primary and secondary lymphoid tissues. *Eur. J. Immunol.* 46: 480-492.
107. Briney, B. S., J. R. Willis, J. A. Finn, B. A. McKinney, and J. E. Crowe, Jr. 2014. Tissue-specific expressed antibody variable gene repertoires. *PLoS one* 9: e100839.
108. Martin, V. G., Y. B. Wu, C. L. Townsend, G. H. Lu, J. S. O'Hare, A. Mozeika, A. C. Coolen, D. Kipling, F. Fraternali, and D. K. Dunn-Walters. 2016. Transitional B Cells in

- Early Human B Cell Development - Time to Revisit the Paradigm? *Front. Immunol.* 7: 546.
109. Turchaninova, M. A., A. Davydov, O. V. Britanova, M. Shugay, V. Bikos, E. S. Egorov, V. I. Kirgizova, E. M. Merzlyak, D. B. Staroverov, D. A. Bolotin, I. Z. Mamedov, M. Izraelson, M. D. Logacheva, O. Kladova, K. Plevova, S. Pospisilova, and D. M. Chudakov. 2016. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat. Protoc.* 11: 1599-1616.
110. Lu, J., T. Panavas, K. Thys, J. Aerssens, M. Naso, J. Fisher, M. Ryczyn, and R. W. Sweet. 2014. IgG variable region and VH CDR3 diversity in unimmunized mice analyzed by massively parallel sequencing. *Mol. Immunol.* 57: 274-283.
111. He, L., D. Sok, P. Azadnia, J. Hsueh, E. Landais, M. Simek, W. C. Koff, P. Poignard, D. R. Burton, and J. Zhu. 2014. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci. Rep.* 4: 6778.
112. Lin, S. G., Z. Ba, Z. Du, Y. Zhang, J. Hu, and F. W. Alt. 2016. Highly sensitive and unbiased approach for elucidating antibody repertoires. *Proc. Natl. Acad. Sci.* 113: 7846-7851.
113. Greiff, V., E. Miho, U. Menzel, and S. T. Reddy. 2015. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. *Trends Immunol.* 36: 738-749.
114. Bolotin, D. A., S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva, and D. M. Chudakov. 2015. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* 12: 380-381.

115. Shugay, M., O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, C. Linnemann, D. Shagin, S. Pospisilova, S. Lukyanov, T. N. Schumacher, and D. M. Chudakov. 2014. Towards error-free profiling of immune repertoires. *Nat. Methods* 11: 653-655.
116. Ye, J., N. Ma, T. L. Madden, and J. M. Ostell. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41: W34-40.
117. Vander Heiden, J. A., G. Yaari, M. Uduman, J. N. Stern, K. C. O'Connor, D. A. Hafler, F. Vigneault, and S. H. Kleinstein. 2014. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics (Oxford, England)* 30: 1930-1932.
118. Lefranc, M. P. 2014. Antibody Informatics: IMGT, the International ImMunoGeneTics Information System. *Microbiol. Spectr.* 2: 1-14.
119. Alamyar, E., P. Duroux, M. P. Lefranc, and V. Giudicelli. 2012. IMGT((R)) 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.* 882: 569-604.
120. Moorhouse, M. J., D. van Zessen, I. J. H. S. Hiltemann, S. Horsman, P. J. van der Spek, M. van der Burg, and A. P. Stubbs. 2014. ImmunoGlobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. *BMC Immunol.* 15: 59.
121. Jspeert, I. H., P. A. van Schouwenburg, D. van Zessen, I. Pico-Knijnenburg, A. P. Stubbs, and M. van der Burg. 2017. Antigen Receptor Galaxy: A User-Friendly, Web-Based Tool

- for Analysis and Visualization of T and B Cell Receptor Repertoire Data. *J. Immunol.* 198: 4156-4165.
122. Huang, L., M. D. Lange, and Z. Zhang. 2014. VH Replacement Footprint Analyzer-I, a Java-Based Computer Program for Analyses of Immunoglobulin Heavy Chain Genes and Potential VH Replacement Products in Human and Mouse. *Front. Immunol.* 5: 40.
  123. Bischof, J., and S. M. Ibrahim. 2016. bcRep: R Package for Comprehensive Analysis of B Cell Receptor Repertoire Data. *PloS one* 11: e0161569.
  124. Schaller, S., J. Weinberger, R. Jimenez-Heredia, M. Danzer, R. Oberbauer, C. Gabriel, and S. M. Winkler. 2015. ImmunExplorer (IMEX): a software framework for diversity and clonality analyses of immunoglobulins and T cell receptors on the basis of IMGT/HighV-QUEST preprocessed NGS data. *BMC Bioinformatics* 16: 252.
  125. Gupta, N. T., J. A. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, and S. H. Kleinstein. 2015. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* 31: 3356-3358.
  126. Zhang, W., Y. Du, Z. Su, C. Wang, X. Zeng, R. Zhang, X. Hong, C. Nie, J. Wu, H. Cao, X. Xu, and X. Liu. 2015. IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis. *Genetics* 201: 459-472.
  127. Cortina-Ceballos, B., E. E. Godoy-Lozano, H. Samano-Sanchez, A. Aguilar-Salgado, C. Velasco-Herrera Mdel, C. Vargas-Chavez, D. Velazquez-Ramirez, G. Romero, J. Moreno, J. Tellez-Sosa, and J. Martinez-Barnette. 2015. Reconstructing and mining the B cell repertoire with ImmunediveRsity. *mAbs* 7: 516-524.
  128. McCoy, C. O., T. Bedford, V. N. Minin, P. Bradley, H. Robins, and F. A. t. Matsen. 2015. Quantifying evolutionary constraints on B-cell affinity maturation. *Phil. Trans. R. Soc.*



- Lon. B.* 370: 20140244.
129. Mirsky, A., L. Kazandjian, and M. Anisimova. 2015. Antibody-specific model of amino acid substitution for immunological inferences from alignments of antibody sequences. *Mol. Biol. Evol.* 32: 806-819.
  130. Barak, M., N. S. Zuckerman, H. Edelman, R. Unger, and R. Mehr. 2008. IgTree: creating Immunoglobulin variable region gene lineage trees. *J. Immunol. Methods* 338: 67-74.
  131. Schramm, C. A., Z. Sheng, Z. Zhang, J. R. Mascola, P. D. Kwong, and L. Shapiro. 2016. SONAR: A High-Throughput Pipeline for Inferring Antibody Ontogenies from Longitudinal Sequencing of B Cell Transcripts. *Front. Immunol.* 7: 372.
  132. Sheng, Z., C. A. Schramm, R. Kong, J. C. Mullikin, J. R. Mascola, P. D. Kwong, and L. Shapiro. 2017. Gene-Specific Substitution Profiles Describe the Types and Frequencies of Amino Acid Changes during Antibody Somatic Hypermutation. *Front. Immunol.* 8: 537.
  133. Briney, B., K. Le, J. Zhu, and D. R. Burton. 2016. Clonify: unseeded antibody lineage assignment from next-generation sequencing data. *Sci. Rep.* 6: 23901.
  134. Gadala-Maria, D., G. Yaari, M. Uduman, and S. H. Kleinstein. 2015. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci.* 112: E862-870.
  135. Corcoran, M. M., G. E. Phad, N. Vazquez Bernat, C. Stahl-Hennig, N. Sumida, M. A. Persson, M. Martin, and G. B. Karlsson Hedestam. 2016. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.* 7: 13642.

136. Kirik, U., L. Greiff, F. Levander, and M. Ohlin. 2017. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Mol. Immunol.* 87: 12-22.
137. Imkeller, K., P. F. Arndt, H. Wardemann, and C. E. Busse. 2016. sciReptor: analysis of single-cell level immunoglobulin repertoires. *BMC Bioinformatics* 17: 67.
138. Toby, I. T., M. K. Levin, E. A. Salinas, S. Christley, S. Bhattacharya, F. Breden, A. Buntzman, B. Corrie, J. Fonner, N. T. Gupta, U. Hershberg, N. Marthandan, A. Rosenfeld, W. Rounds, F. Rubelt, W. Scarborough, J. K. Scott, M. Uduman, J. A. Vander Heiden, R. H. Scheuermann, N. Monson, S. H. Kleinstein, and L. G. Cowell. 2016. VDJML: a file format with tools for capturing the results of inferring immune receptor rearrangements. *BMC Bioinformatics* 17: 333.
139. Safonova, Y., A. Lapidus, and J. Lill. 2015. IgSimulator: a versatile immunosequencing simulator. *Bioinformatics (Oxford, England)* 31: 3213-3215.
140. Luo, S., J. A. Yu, and Y. S. Song. 2016. Estimating Copy Number and Allelic Variation at the Immunoglobulin Heavy Chain Locus Using Short Reads. *PLoS Comput. Biol.* 12: e1005117.
141. Rettig, T. A., C. Ward, M. J. Pecaut, and S. K. Chapes. 2017. Validation of Methods to Assess the Immunoglobulin Gene Repertoire in Tissues Obtained from Mice on the International Space Station. *Gavit. Space Res.* 5:2-23.
142. Kaplinsky, J., A. Li, A. Sun, M. Coffre, S. B. Koralov, and R. Arnaout. 2014. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc. Natl. Acad. Sci.* 111: E2622-2629.

143. Reddy, S. T., X. Ge, A. E. Miklos, R. A. Hughes, S. H. Kang, K. H. Hoi, C. Chrysostomou, S. P. Hunicke-Smith, B. L. Iverson, P. W. Tucker, A. D. Ellington, and G. Georgiou. 2010. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature Biotechnol.* 28: 965-969.
144. Choi, N. M., S. Loguercio, J. Verma-Gaur, S. C. Degner, A. Torkamani, A. I. Su, E. M. Oltz, M. Artyomov, and A. J. Feeney. 2013. Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. *J. Immunol.* 191: 2393-2402.
145. Yang, Y., C. Wang, Q. Yang, A. B. Kantor, H. Chu, E. E. Ghosn, G. Qin, S. K. Mazmanian, J. Han, and L. A. Herzenberg. 2015. Distinct mechanisms define murine B cell lineage immunoglobulin heavy chain (IgH) repertoires. *eLife* 4: e09083.
146. Holodick, N. E., L. Zeumer, T. L. Rothstein, and L. Morel. 2016. Expansion of B-1a Cells with Germline Heavy Chain Sequence in Lupus Mice. *Front. Immunol.* 7: 108.
147. Aoki-Ota, M., A. Torkamani, T. Ota, N. Schork, and D. Nemazee. 2012. Skewed primary Ighkappa repertoire and V-J joining in C57BL/6 mice: implications for recombination accessibility and receptor editing. *J. Immunol.* 188: 2305-2315.
148. Collins, A. M., Y. Wang, K. M. Roskin, C. P. Marquis, and K. J. Jackson. 2015. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Phil. Trans. R. Soc. Lon. B* 370: 20140236.
149. Abolins, S., E. C. King, L. Lazarou, L. Weldon, L. Hughes, P. Drescher, J. G. Raynes, J. C. R. Hafalla, M. E. Viney, and E. M. Riley. 2017. The comparative immunology of wild and laboratory mice, *Mus musculus domesticus*. *Nat. Commun.* 8: 14811.
150. Kono, N., L. Sun, H. Toh, T. Shimizu, H. Xue, O. Numata, M. Ato, K. Ohnishi, and S.

- Itamura. 2017. Deciphering antigen-responding antibody repertoires by using next-generation sequencing and confirming them through antibody-gene synthesis. *Biochem. Biophys. Res. Commun.* 487: 300-306.
151. Chen, H. S., S. C. Hou, J. W. Jian, K. S. Goh, S. T. Shen, Y. C. Lee, J. J. You, H. P. Peng, W. C. Kuo, S. T. Chen, M. C. Peng, A. H. Wang, C. M. Yu, I. C. Chen, C. P. Tung, T. H. Chen, K. Ping Chiu, C. Ma, C. Yuan Wu, S. W. Lin, and A. S. Yang. 2015. Predominant structural configuration of natural antibody repertoires enables potent antibody responses against protein antigens. *Sc. Rep.* 5: 12411.
152. Greiff, V., U. Menzel, E. Miho, C. Weber, R. Riedel, S. Cook, A. Valai, T. Lopes, A. Radbruch, T. H. Winkler, and S. T. Reddy. 2017. Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep.* 19: 1467-1478.
153. Jspeert, I. H., P. A. van Schouwenburg, D. van Zessen, I. Pico-Knijnenburg, G. J. Driessen, A. P. Stubbs, and M. van der Burg. 2016. Evaluation of the Antigen-Experienced B-Cell Receptor Repertoire in Healthy Children and Adults. *Front. Immunol.* 7: 410.
154. Wu, Y. C., D. Kipling, H. S. Leong, V. Martin, A. A. Ademokun, and D. K. Dunn-Walters. 2010. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* 116: 1070-1078.
155. Wu, Y. C., D. Kipling, and D. K. Dunn-Walters. 2012. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. *Front. Immunol.* 3: 193.
156. Wang, C., Y. Liu, L. T. Xu, K. J. Jackson, K. M. Roskin, T. D. Pham, J. Laserson, E. L. Marshall, K. Seo, J. Y. Lee, D. Furman, D. Koller, C. L. Dekker, M. M. Davis, A. Z. Fire,

- and S. D. Boyd. 2014. Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires. *J. Immunol.* 192: 603-611.
157. Mroczek, E. S., G. C. Ippolito, T. Rogosch, K. H. Hoi, T. A. Hwangpo, M. G. Brand, Y. Zhuang, C. R. Liu, D. A. Schneider, M. Zemlin, E. E. Brown, G. Georgiou, and H. W. Schroeder, Jr. 2014. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Front. Immunol.* 5: 96.
158. Arnaout, R., W. Lee, P. Cahill, T. Honan, T. Sparrow, M. Weiland, C. Nusbaum, K. Rajewsky, and S. B. Koralov. 2011. High-resolution description of antibody heavy-chain repertoires in humans. *PloS one* 6: e22365.
159. Benichou, J., J. Glanville, E. T. Prak, R. Azran, T. C. Kuo, J. Pons, C. Desmarais, L. Tsaban, and Y. Louzoun. 2013. The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. *J. Immunol.* 190: 5567-5577.
160. Larimore, K., M. W. McCormick, H. S. Robins, and P. D. Greenberg. 2012. Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* 189: 3221-3230.
161. Briney, B. S., J. R. Willis, M. D. Hicar, J. W. Thomas, 2nd, and J. E. Crowe, Jr. 2012. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology* 137: 56-64.
162. Briney, B. S., J. R. Willis, and J. E. Crowe, Jr. 2012. Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Gene Immun.* 13: 523-529.

163. Galson, J. D., J. Trück, A. Fowler, M. Münz, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. In-Depth Assessment of Within-Individual and Inter-Individual Variation in the B Cell Receptor Repertoire. *Front. Immunol.* 6.
164. Hoi, K. H., and G. C. Ippolito. 2013. Intrinsic bias and public rearrangements in the human immunoglobulin Vlambda light chain repertoire. *Genes Immun.* 14: 271-276.
165. Jackson, K. J., Y. Wang, B. A. Gaeta, W. Pomat, P. Siba, J. Rimmer, W. A. Sewell, and A. M. Collins. 2012. Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. *Immunogenetics* 64: 3-14.
166. Boyd, S. D., B. A. Gaeta, K. J. Jackson, A. Z. Fire, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, and A. M. Collins. 2010. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* 184: 6986-6992.
167. Wang, Y., K. J. Jackson, B. Gaeta, W. Pomat, P. Siba, W. A. Sewell, and A. M. Collins. 2011. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* 63: 259-265.
168. Snir, O., L. Mesin, M. Gidoni, K. E. Lundin, G. Yaari, and L. M. Sollid. 2015. Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing. *J. Immunol.* 194: 5703-5712.
169. Tan, Y. C., S. Kongpachith, L. K. Blum, C. H. Ju, L. J. Lahey, D. R. Lu, X. Cai, C. A. Wagner, T. M. Lindstrom, J. Sokolove, and W. H. Robinson. 2014. Barcode-enabled sequencing of plasmablast antibody repertoires in rheumatoid arthritis. *Arthritis Rheumatol.* 66: 2706-2715.

170. Stern, J. N., G. Yaari, J. A. Vander Heiden, G. Church, W. F. Donahue, R. Q. Hintzen, A. J. Huttner, J. D. Laman, R. M. Nagra, A. Nylander, D. Pitt, S. Ramanan, B. A. Siddiqui, F. Vigneault, S. H. Kleinstein, D. A. Hafler, and K. C. O'Connor. 2014. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.* 6: 248ra107.
171. von Budingen, H. C., T. C. Kuo, M. Sirota, C. J. van Belle, L. Apeltsin, J. Glanville, B. A. Cree, P. A. Gourraud, A. Schwartzburg, G. Huerta, D. Telman, P. D. Sundar, T. Casey, D. R. Cox, and S. L. Hauser. 2012. B cell exchange across the blood-brain barrier in multiple sclerosis. *J. Clin. Invest.* 122: 4533-4543.
172. Xiao, M., P. Prabakaran, W. Chen, B. Kessing, and D. S. Dimitrov. 2013. Deep sequencing and Circos analyses of antibody libraries reveal antigen-driven selection of Ig VH genes during HIV-1 infection. *Exp. Mol. Pathol.* 95: 357-363.
173. Jiang, Y., K. Nie, D. Redmond, A. M. Melnick, W. Tam, and O. Elemento. 2015. VDJ-Seq: Deep Sequencing Analysis of Rearranged Immunoglobulin Heavy Chain Gene to Reveal Clonal Evolution Patterns of B Cell Lymphoma. *JoVE*: e53215.
174. Logan, A. C., H. Gao, C. Wang, B. Sahaf, C. D. Jones, E. L. Marshall, I. Buno, R. Armstrong, A. Z. Fire, K. I. Weinberg, M. Mindrinos, J. L. Zehnder, S. D. Boyd, W. Xiao, R. W. Davis, and D. B. Miklos. 2011. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc. Natl. Acad. Sci.* 108: 21194-21199.
175. Bashford-Rogers, R. J., A. L. Palser, B. J. Huntly, R. Rance, G. S. Vassiliou, G. A. Follows, and P. Kellam. 2013. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res.* 23: 1874-1884.

176. Rene, C., N. Prat, A. Thuizat, M. Broctawik, O. Avinens, and J. F. Eliaou. 2014. Comprehensive characterization of immunoglobulin gene rearrangements in patients with chronic lymphocytic leukaemia. *J. Cell. Mol. Med.* 18: 979-990.
177. Tschumper, R. C., Y. W. Asmann, A. Hossain, P. M. Huddleston, X. Wu, A. Dispenzieri, B. W. Eckloff, and D. F. Jelinek. 2012. Comprehensive assessment of potential multiple myeloma immunoglobulin heavy chain V-D-J intraclonal variation using massively parallel pyrosequencing. *Oncotarget* 3: 502-513.
178. Beausang, J. F., H. C. Fan, R. Sit, M. U. Hutchins, K. Jirage, R. Curtis, E. Hutchins, S. R. Quake, and J. M. Yabu. 2017. B cell repertoires in HLA-sensitized kidney transplant candidates undergoing desensitization therapy. *J. Transl. Med.* 15: 9.
179. Levin, M., J. J. King, J. Glanville, K. J. Jackson, T. J. Looney, R. A. Hoh, A. Mari, M. Andersson, L. Greiff, A. Z. Fire, S. D. Boyd, and M. Ohlin. 2016. Persistence and evolution of allergen-specific IgE repertoires during subcutaneous specific immunotherapy. *J. Allergy Clin. Immunol.* 137: 1535-1544.
180. Lee, J., D. R. Boutz, V. Chromikova, M. G. Joyce, C. Vollmers, K. Leung, A. P. Horton, B. J. DeKosky, C. H. Lee, J. J. Lavinder, E. M. Murrin, C. Chrysostomou, K. H. Hoi, Y. Tsybovsky, P. V. Thomas, A. Druz, B. Zhang, Y. Zhang, L. Wang, W. P. Kong, D. Park, L. I. Popova, C. L. Dekker, M. M. Davis, C. E. Carter, T. M. Ross, A. D. Ellington, P. C. Wilson, E. M. Marcotte, J. R. Mascola, G. C. Ippolito, F. Krammer, S. R. Quake, P. D. Kwong, and G. Georgiou. 2016. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat. Med.* 22: 1456-1464.
181. Laserson, U., F. Vigneault, D. Gadala-Maria, G. Yaari, M. Uduman, J. A. Vander Heiden, W. Kelton, S. Taek Jung, Y. Liu, J. Laserson, R. Chari, J. H. Lee, I. Bachelet, B. Hickey,



- E. Lieberman-Aiden, B. Hanczaruk, B. B. Simen, M. Egholm, D. Koller, G. Georgiou, S. H. Kleinstein, and G. M. Church. 2014. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci.* 111: 4928-4933.
182. Truck, J., M. N. Ramasamy, J. D. Galson, R. Rance, J. Parkhill, G. Lunter, A. J. Pollard, and D. F. Kelly. 2015. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J. Immunol.* 194: 252-261.
183. Horns, F., C. Vollmers, D. Croote, S. F. Mackey, G. E. Swan, C. L. Dekker, M. M. Davis, and S. R. Quake. 2016. Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *eLife* 5.
184. Lavinder, J. J., Y. Wine, C. Giesecke, G. C. Ippolito, A. P. Horton, O. I. Lungu, K. H. Hoi, B. J. DeKosky, E. M. Murrin, M. M. Wirth, A. D. Ellington, T. Dorner, E. M. Marcotte, D. R. Boutz, and G. Georgiou. 2014. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc. Natl. Acad. Sci.* 111: 2259-2264.
185. Halliley, J. L., C. M. Tipton, J. Liesveld, A. F. Rosenberg, J. Darce, I. V. Gregoret, L. Popova, D. Kaminiski, C. F. Fucile, I. Albizua, S. Kyu, K. Y. Chiang, K. T. Bradley, R. Burack, M. Slifka, E. Hammarlund, H. Wu, L. Zhao, E. E. Walsh, A. R. Falsey, T. D. Randall, W. C. Cheung, I. Sanz, and F. E. Lee. 2015. Long-Lived Plasma Cells Are Contained within the CD19(-)CD38(hi)CD138(+) Subset in Human Bone Marrow. *Immunity* 43: 132-145.
186. Galson, J. D., E. A. Clutterbuck, J. Truck, M. N. Ramasamy, M. Munz, A. Fowler, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. *Immunol. Cell Biol.* 93: 885-895.

187. Liu, J., R. Li, K. Liu, L. Li, X. Zai, X. Chi, L. Fu, J. Xu, and W. Chen. 2016. Identification of antigen-specific human monoclonal antibodies using high-throughput sequencing of the antibody repertoire. *Biochem. Biophys. Res. Commun.* 473: 23-28.
188. Galson, J. D., J. Truck, E. A. Clutterbuck, A. Fowler, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. 2016. B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome Med.* 8: 68.
189. Vollmers, C., R. V. Sit, J. A. Weinstein, C. L. Dekker, and S. R. Quake. 2013. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci.* 110: 13463-13468.
190. Greiff, V., P. Bhat, S. C. Cook, U. Menzel, W. Kang, and S. T. Reddy. 2015. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* 7: 49.
191. Hou, D., T. Ying, L. Wang, C. Chen, S. Lu, Q. Wang, E. Seeley, J. Xu, X. Xi, T. Li, J. Liu, X. Tang, Z. Zhang, J. Zhou, C. Bai, C. Wang, M. Byrne-Steele, J. Qu, J. Han, and Y. Song. 2016. Immune Repertoire Diversity Correlated with Mortality in Avian Influenza A (H7N9) Virus Infected Patients. *Sci. Rep.* 6: 33843.
192. Rappuoli, R., M. J. Bottomley, U. D'Oro, O. Finco, and E. De Gregorio. 2016. Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design. *J. Exp. Med.* 213: 469-481.
193. Blaber, E., H. Marcal, and B. P. Burns. 2010. Bioastronautics: the influence of microgravity on astronaut health. *Astrobiology* 10: 463-473.
194. Berry, C. A. 1970. Summary of medical experience in the Apollo 7 through 11 manned spaceflights. *Aerospace Med.* 41: 500-519.

195. Crucian, B., R. Stowe, S. Mehta, P. Uchakin, H. Quiariarte, D. Pierson, and C. Sams. 2013. Immune system dysregulation occurs during short duration spaceflight on board the space shuttle. *J. Clin. Immunol.* 33: 456-465.
196. Grigoriev, A. I., S. A. Bugrov, V. V. Bogomolov, A. D. Egorov, V. V. Polyakov, I. K. Tarasov, and E. B. Shulzhenko. 1993. Main medical results of extended flights on space station Mir in 1986-1990. *Acta Astronautica* 29: 581-585.
197. Taylor, G. R., L. S. Neale, and J. R. Dardano. 1986. Immunological analyses of U.S. Space Shuttle crewmembers. *Aviat. Space Environ. Med.* 57: 213-217.
198. Taylor, G. R. 1993. Immune changes during short-duration missions. *J. Leukoc. Biol.* 54: 202-208.
199. Crucian, B., R. J. Simpson, S. Mehta, R. Stowe, A. Chouker, S. A. Hwang, J. K. Actor, A. P. Salam, D. Pierson, and C. Sams. 2014. Terrestrial stress analogs for spaceflight associated immune system dysregulation. *Brain Behav. Immun.* 39: 23-32.
200. Stein, T. P., and M. D. Schluter. 1994. Excretion of IL-6 by astronauts during spaceflight. *Am. J. Physiol.* 266: E448-452.
201. Globus, R. K., and E. Morey-Holton. 2016. Hindlimb unloading: rodent analog for microgravity. *J. Appl. Physiol.* 120: 1196-1206.
202. Chapes, S. K., A. M. Mastro, G. Sonnenfeld, and W. D. Berry. 1993. Antiorthostatic suspension as a model for the effects of spaceflight on the immune system. *J. Leukoc. Biol.* 54: 227-235.
203. Pecaut, M. J., S. J. Simske, and M. Fleshner. 2000. Spaceflight induces changes in splenocyte subpopulations: effectiveness of ground-based models. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 279: R2072-2078.

204. Stowe, R. P., D. L. Yetman, W. F. Storm, C. F. Sams, and D. L. Pierson. 2008. Neuroendocrine and immune responses to 16-day bed rest with realistic launch and landing G profiles. *Aviat. Space Environ. Med.* 79: 117-122.
205. Tauber, S., S. Hauschild, K. Paulsen, A. Gutewort, C. Raig, E. Hurlimann, J. Biskup, C. Philpot, H. Lier, F. Engelmann, A. Pantaleo, A. Cogoli, P. Pippia, L. E. Layer, C. S. Thiel, and O. Ullrich. 2015. Signal transduction in primary human T lymphocytes in altered gravity during parabolic flight and clinostat experiments. *Cell. Physiol. Biochem.* 35: 1034-1051.
206. Kita, M., T. Yamamoto, J. Imanishi, and A. Fuse. 2004. Influence of gravity changes induced by parabolic flight on cytokine production in mouse spleen. *J. Gravit. Physiol.* 11: P67-68.
207. Nickerson, C. A., C. M. Ott, J. W. Wilson, R. Ramamurthy, C. L. LeBlanc, K. Honer zu Bentrup, T. Hammond, and D. L. Pierson. 2003. Low-shear modeled microgravity: a global environmental regulatory signal affecting bacterial gene expression, physiology, and pathogenesis. *J. Microbiol. Methods* 54: 1-11.
208. Gridley, D. S., R. Dutta-Roy, M. L. Andres, G. A. Nelson, and M. J. Pecaut. 2006. *Radiat. Res.* 165: 78-87.
209. Gridley, D. S., and M. J. Pecaut. 2016. Changes in the distribution and function of leukocytes after whole-body iron ion irradiation. *J. Radiat. Res.* 57: 477-491.
210. Li, M., V. Holmes, Y. Zhou, H. Ni, J. K. Sanzari, A. R. Kennedy, and D. Weissman. 2014. Hindlimb suspension and SPE-like radiation impairs clearance of bacterial infections. *PloS one* 9: e85665.

211. Gaignier, F., V. Schenten, M. De Carvalho Bittencourt, G. Gauquelin-Koch, J. P. Frippiat, and C. Legrand-Frossi. 2014. Three weeks of murine hindlimb unloading induces shifts from B to T and from th to tc splenic lymphocytes in absence of stress and differentially reduces cell-specific mitogenic responses. *PloS one* 9: e92664.
212. Grove, D. S., S. A. Pishak, and A. M. Mastro. 1995. The effect of a 10-day space flight on the function, phenotype, and adhesion molecule expression of splenocytes and lymph node lymphocytes. *Exp. Cell Res.* 219: 102-109.
213. Murphy, K., P. Travers, M. Walport, and C. Janeway. 2012. *Janeway's Immunobiology* Garland Science, New York.
214. Allebban, Z., A. T. Ichiki, L. A. Gibson, J. B. Jones, C. C. Congdon, and R. D. Lange. 1994. Effects of spaceflight on the number of rat peripheral blood leukocytes and lymphocyte subsets. *J. Leukoc. Biol.* 55: 209-213.
215. Ichiki, A. T., L. A. Gibson, T. L. Jago, K. M. Strickland, D. L. Johnson, R. D. Lange, and Z. Allebban. 1996. Effects of spaceflight on rat peripheral blood leukocytes and bone marrow progenitor cells. *J. Leukoc. Biol.* 60: 37-43.
216. Chapes, S. K., S. J. Simske, G. Sonnenfeld, E. S. Miller, and R. J. Zimmerman. 1999. Effects of spaceflight and PEG-IL-2 on rat physiological and immunological responses. *J. Appl. Physiol.* 86: 2065-2076.
217. Gridley, D. S., J. M. Slater, X. Luo-Owen, A. Rizvi, S. K. Chapes, L. S. Stodieck, V. L. Ferguson, and M. J. Pecaute. 2009. Spaceflight effects on T lymphocyte distribution, function and gene expression. *J. Appl. Physiol.* 106: 194-202.

218. Gridley, D. S., X. W. Mao, L. S. Stodieck, V. L. Ferguson, T. A. Bateman, M. Moldovan, C. E. Cunningham, T. A. Jones, J. M. Slater, and M. J. Pecaut. 2013. Changes in mouse thymus and spleen after return from the STS-135 mission in space. *PLoS one* 8: e75097.
219. Pecaut, M. J., G. A. Nelson, L. L. Peters, P. J. Kostenuik, T. A. Bateman, S. Morony, L. S. Stodieck, D. L. Lacey, S. J. Simske, and D. S. Gridley. 2003. Genetic models in applied physiology: selected contribution: effects of spaceflight on immunity in the C57BL/6 mouse. I. Immune population distributions. *J. Appl. Physiol.* 94: 2085-2094.
220. Sonnenfeld, G., A. D. Mandel, I. V. Konstantinova, G. R. Taylor, W. D. Berry, S. R. Wellhausen, A. T. Lesnyak, and B. B. Fuchs. 1990. Effects of spaceflight on levels and activity of immune cells. *Aviat. Space Environ. Med.* 61: 648-653.
221. Sonnenfeld, G., A. D. Mandel, I. V. Konstantinova, W. D. Berry, G. R. Taylor, A. T. Lesnyak, B. B. Fuchs, and A. L. Rakhmievich. 1992. Spaceflight alters immune cell function and distribution. *J. Appl. Physiol.* 73: 191s-195s.
222. Wei, L. X., J. N. Zhou, A. I. Roberts, and Y. F. Shi. 2003. Lymphocyte reduction induced by hindlimb unloading: distinct mechanisms in the spleen and thymus. *Cell Res.* 13: 465-471.
223. Baqai, F. P., D. S. Gridley, J. M. Slater, X. Luo-Owen, L. S. Stodieck, V. Ferguson, S. K. Chapes, and M. J. Pecaut. 2009. Effects of spaceflight on innate immune function and antioxidant gene expression. *J. Appl. Physiol. (Bethesda, Md. : 1985)* 106: 1935-1942.
224. Chapes, S. K., S. J. Simske, A. D. Forsman, T. A. Bateman, and R. J. Zimmerman. 1999. Effects of space flight and IGF-1 on immune function. *Adv. Space Res.* 23: 1955-1964.

225. Congdon, C. C., Z. Allebban, L. A. Gibson, A. Kaplansky, K. M. Strickland, T. L. Jago, D. L. Johnson, R. D. Lange, and A. T. Ichiki. 1996. Lymphatic tissue changes in rats flown on Spacelab Life Sciences-2. *J. Appl. Physiol.* 81: 172-177.
226. Durnova, G. N., A. S. Kaplansky, and V. V. Portugalov. 1976. Effect of a 22-day space flight on the lymphoid organs of rats. *Aviat. Space Environ. Med.* 47: 588-591.
227. Armstrong, J. W., K. A. Nelson, S. J. Simske, M. W. Luttgess, J. J. Iandolo, and S. K. Chapes. 1993. Skeletal unloading causes organ-specific changes in immune cell responses. *J. Appl. Physiol.* 75: 2734-2739.
228. Nash, P. V., B. A. Bour, and A. M. Mastro. 1991. Effect of hindlimb suspension simulation of microgravity on in vitro immunological responses. *Exp. Cell Res.* 195: 353-360.
229. Lebsack, T. W., V. Fa, C. C. Woods, R. Gruener, A. M. Manziello, M. J. Pecaut, D. S. Gridley, L. S. Stodieck, V. L. Ferguson, and D. Deluca. 2010. Microarray analysis of spaceflown murine thymus tissue reveals changes in gene expression regulating stress and glucocorticoid receptors. *J. Cell. Biochem.* 110: 372-381.
230. O'Donnell, P. M., J. M. Orshal, D. Sen, G. Sonnenfeld, and H. O. Aviles. 2009. Effects of exposure of mice to hindlimb unloading on leukocyte subsets and sympathetic nervous system activity. *Stress* 12: 82-88.
231. Battista, N., M. A. Meloni, M. Bari, N. Mastrangelo, G. Galleri, C. Rapino, E. Dainese, A. F. Agro, P. Pippia, and M. Maccarrone. 2012. 5-Lipoxygenase-dependent apoptosis of human lymphocytes in the International Space Station: data from the ROALD experiment. *FASEB J.* 26: 1791-1798.

232. Ortega, M. T., M. J. Pecaut, D. S. Gridley, L. S. Stodieck, V. Ferguson, and S. K. Chapes. 2009. Shifts in bone marrow cell phenotypes caused by spaceflight. *J. Appl. Physiol.* 106: 548-555.
233. Lescale, C., V. Schenten, D. Djeghloul, M. Bennabi, F. Gagnier, K. Vandamme, C. Strazielle, I. Kuzniak, H. Petite, C. Dosquet, J. P. Fripiat, and M. Goodhardt. 2015. Hind limb unloading, a model of spaceflight conditions, leads to decreased B lymphopoiesis similar to aging. *FASEB J.* 29: 455-463.
234. Chang, T. T., I. Walther, C. F. Li, J. Boonyaratanakornkit, G. Galleri, M. A. Meloni, P. Pippia, A. Cogoli, and M. Hughes-Fulford. 2012. The Rel/NF-kappaB pathway and transcription of immediate early genes in T cell activation are inhibited by microgravity. *J. Leukoc. Biol.* 92: 1133-1145.
235. Hwang, S. A., B. Crucian, C. Sams, and J. K. Actor. 2015. Post-Spaceflight (STS-135) Mouse Splenocytes Demonstrate Altered Activation Properties and Surface Molecule Expression. *PloS one* 10: e0124380.
236. Martinez, E. M., M. C. Yoshida, T. L. Candelario, and M. Hughes-Fulford. 2015. Spaceflight and simulated microgravity cause a significant reduction of key gene expression in early T-cell activation. *Am. J. Physiol.* 308: R480-488.
237. Sanzari, J. K., A. L. Romero-Weaver, G. James, G. Krigsfeld, L. Lin, E. S. Diffenderfer, and A. R. Kennedy. 2013. Leukocyte activity is altered in a ground based murine model of microgravity and proton radiation exposure. *PloS one* 8: e71757.
238. Cooper, D., M. W. Pride, E. L. Brown, D. Risin, and N. R. Pellis. 2001. Suppression of antigen-specific lymphocyte activation in modeled microgravity. *In Vitro Cell. Dev. Biol. Anim.* 37: 63-65.



239. Lesnyak, A. T., G. Sonnenfeld, M. P. Rykova, D. O. Meshkov, A. Mastro, and I. Konstantinova. 1993. Immune changes in test animals during spaceflight. *J. Leukoc. Biol.* 54: 214-226.
240. Lesnyak, A., G. Sonnenfeld, L. Avery, I. Konstantinova, M. Rykova, D. Meshkov, and T. Orlova. 1996. Effect of SLS-2 spaceflight on immunologic parameters of rats. *J. Appl. Physiol.* 81: 178-182.
241. Nash, P. V., and A. M. Mastro. 1992. Variable lymphocyte responses in rats after space flight. *Exp. Cell. Res.* 202: 125-131.
242. Nash, P. V., I. V. Konstantinova, B. B. Fuchs, A. L. Rakhmievich, A. T. Lesnyak, and A. M. Mastro. 1992. Effect of spaceflight on lymphocyte proliferation and interleukin-2 production. *J. Appl. Physiol.* 73: 186s-190s.
243. Sonnenfeld, G., M. Foster, D. Morton, F. Bailliard, N. A. Fowler, A. M. Hakenewerth, R. Bates, and E. S. Miller, Jr. 1998. Spaceflight and development of immune responses. *J. Appl. Physiol.* 85: 1429-1433.
244. Gould, C. L., M. Lyte, J. Williams, A. D. Mandel, and G. Sonnenfeld. 1987. Inhibited interferon-gamma but normal interleukin-3 production from rats flown on the space shuttle. *Aviat. Space Environ. Med.* 58: 983-986.
245. Miller, E. S., D. A. Koebel, and G. Sonnenfeld. 1995. Influence of spaceflight on the production of interleukin-3 and interleukin-6 by rat spleen and thymus cells. *J. Appl. Physiol.* 78: 810-813.
246. Crucian, B. E., S. R. Zwart, S. Mehta, P. Uchakin, H. D. Quiariarte, D. Pierson, C. F. Sams, and S. M. Smith. 2014. Plasma cytokine concentrations indicate that in vivo hormonal

- regulation of immunity is altered during long-duration spaceflight. *J. Interferon Cytokine Res.* 34: 778-786.
247. Fitzgerald, W., S. Chen, C. Walz, J. Zimmerberg, L. Margolis, and J. C. Grivel. 2009. Immune suppression of human lymphoid tissues and cells in rotating suspension culture and onboard the International Space Station. *In Vitro Cell. Dev. Biol. Anim.* 45: 622-632.
248. Pellis, N. R., T. J. Goodwin, D. Risin, B. W. McIntyre, R. P. Pizzini, D. Cooper, T. L. Baker, and G. F. Spaulding. 1997. Changes in gravity inhibit lymphocyte locomotion through type I collagen. *In Vitro Cell. Dev. Biol. Anim.* 33: 398-405.
249. Sundaresan, A., D. Risin, and N. R. Pellis. 2002. Loss of signal transduction and inhibition of lymphocyte locomotion in a ground-based model of microgravity. *In Vitro Cell. Dev. Biol. Anim.* 38: 118-122.
250. Ward, N. E., N. R. Pellis, S. A. Risin, and D. Risin. 2006. Gene expression alterations in activated human T-cells induced by modeled microgravity. *J. Cell. Biochem.* 99: 1187-1202.
251. Hughes-Fulford, M., T. T. Chang, E. M. Martinez, and C. F. Li. 2015. Spaceflight alters expression of microRNA during T-cell activation. *FASEB J.* 29: 4893-4900.
252. Ghislin, S., N. Ouzren-Zarhloul, S. Kaminski, and J. P. Frippiat. 2015. Hypergravity exposure during gestation modifies the TCRbeta repertoire of newborn mice. *Sci. Rep.* 5: 9318.
253. Boxio, R., C. Dournon, and J. P. Frippiat. 2005. Effects of a long-term spaceflight on immunoglobulin heavy chains of the urodele amphibian *Pleurodeles waltl*. *J Appl. Physiol.* 98: 905-910.

254. Bascove, M., C. Huin-Schohn, N. Gueguinou, E. Tschirhart, and J. P. Frippiat. 2009. Spaceflight-associated changes in immunoglobulin VH gene expression in the amphibian *Pleurodeles waltl*. *FASEB J.* 23: 1607-1615.
255. Huin-Schohn, C., N. Gueguinou, V. Schenten, M. Bascove, G. G. Koch, S. Baatout, E. Tschirhart, and J. P. Frippiat. 2013. Gravity changes during animal development affect IgM heavy-chain transcription and probably lymphopoiesis. *FASEB J.* 27: 333-341.

## **Chapter 2 - Validation of Methods to Assess the Immunoglobulin Gene Repertoire in Tissues Obtained from Mice on the International Space Station**

### **Abstract**

Spaceflight is known to affect immune cell populations. In particular, splenic B-cell numbers decrease during spaceflight and in ground-based physiological models. Although antibody isotype changes have been assessed during and after space flight, an extensive characterization of the impact of spaceflight on antibody composition has not been conducted in mice. High-throughput sequencing and bioinformatic tools are now available to assess antibody repertoires. We can now identify immunoglobulin gene segment usage, junctional regions, and modifications that contribute to specificity and diversity. Due to limitations on the International Space Station, alternate sample collection and storage methods must be employed. Our group compared Illumina MiSeq sequencing data from multiple sample preparation methods in normal C57Bl/6J mice to validate that sample preparation and storage would not bias the outcome of antibody repertoire characterization. In this report, we also compared sequencing techniques and a bioinformatic workflow on the data output when we assessed the IgH and Igκ variable gene usage. Our bioinformatic workflow has been optimized for Illumina HiSeq and MiSeq datasets, and is specifically designed to reduce bias, capture the most information from Ig sequences, and produce a data set that provides other data mining options.

This chapter has been published in *Gravitational and Space Research*, see Appendix A.1 for a statement of copyright release.

## Introduction

For B-cell development and specificity, there are a large number of heavy chain (IgH) and kappa light chain (Igκ) gene segments that are used to produce the immunoglobulin (Ig) receptor population repertoire (1). This antibody repertoire is quite large and the possible specificities can theoretically exceed the number of actual antibody molecules in the host (2). In antibodies, the antigen binding region is formed by six complementarity determining regions (CDRs) that loop out from the V region backbone formed by two beta-pleated sheets (3, 4). The germline V gene segment repertoire is necessary for host responses to pathogens and CDR1 and CDR2 are completely encoded for by variable (V region) gene segments (5). Therefore, knowing which V-gene segments are utilized is fundamental to understanding B-cell specificity and the development of effective immune responses. The CDR3 of both the heavy and light chains are highly variable due to their unique generation. During the creation of each Ig sequence, partially random splicing between V-, D- (heavy chain), and J-gene segments occurs and random base insertions occur, called “n-nucleotide” additions. One hypothesis is that V-gene segments have been maintained in the genome because of their importance in binding specific pathogens and provide essential host defense functions. However, CDR3 may be important because it is highly variable. Its role as a highly diverse and variable region is what provides the essential key to determining antigen specificity (6).

The spaceflight environment can impact many parameters critical to the host immune response. In multiple species spaceflight affects the total body, thymus and spleen mass (7-18), circulating corticosterone (11, 14, 19-26), mitogen-induced proliferation, cytokine production and reactivity (14, 18, 19, 22, 27-42), and lymphocyte subpopulation distributions (10, 30, 41, 43-46). Clearly there are broad physiological impacts that affect many systems. The broader implications

of this have been seen in space and ground-based models as detailed by Sonnenfeld and Crucian (47, 48).

B cells are among the immune components that are affected by spaceflight. The number of B cells in the spleen was reduced in mice flown on the space shuttle flight, STS-118 (49). The percentage of B cells in the bone marrow and spleen was also reduced in mice subjected to hindlimb unloading (50, 51). When rats were injected with sheep red blood cells 8 days prior to an 18.5-day COSMOS flight there were lower IgG concentrations compared to both immunized and non-immunized ground controls after landing (19). IgM production was virtually eliminated in lymphocytes stimulated *in vitro* with pokeweed mitogen (PWM) on the International Space Station (ISS) (52). In this study, cells were activated on Earth, frozen down, and then put back into suspension in space where IgM secretion was significantly lower than similarly treated ground-based controls (52). In a study of long-term ISS crewmember in-flight and post flight plasma samples, no significant changes to adaptive immunity or cytokine profiles were detected (53). However, an assessment of peripheral blood revealed changes in the distribution of B cells and a reduction in T-cell function following mitogen stimulation (47).

Our group is interested in the impact of spaceflight on B-cell immunoglobulin gene usage. Next Generation Sequencing (NGS) now allows us to analyze the repertoire of Ig gene segments that are used in the assembly of immunoglobulins that are transcribed by B cells and that are present in the host. NGS allows the determination of V-, D- and J-gene segment usage, CDR3 assembly and the assessment of mutations that occur in response to immune challenge. In space, there exist certain limitations associated with tissue collection and storage methods traditionally used on the ground. In preparation for sending mice to the ISS, our group needed to validate our procedures and our ability to obtain high-quality RNA that could be used for NGS for the

assessment of Ig gene usage. Our group also sought to validate the usage of RNA extracted from whole tissue collected and stored with these limitations in mind. It was also necessary to develop a workflow that would facilitate the analysis of large amounts of data that would be generated during this project. In this manuscript, we describe the development of our workflow and validation of mouse NGS processing protocols for use with space flight experiments.

## Materials and Methods

### Sample Preparation and RNA Extraction

Spleens were removed from four 11-week-old, specific pathogen-free, female C57BL/6J mice housed in the laboratory animal care and services vivarium at Kansas State University. One-half of each spleen was homogenized with a 70  $\mu$ M sieve to generate a single-cell suspension designated, “**cells**”. Spleen cells were pelleted at 350 x g and were resuspended in 5 mL of ice-cold ACK lysing buffer (155mM NH<sub>4</sub>Cl, 10mM KHC0<sub>3</sub>, 0.1mM EDTA) to lyse red blood cells. After five minutes, 10 mL of ice-cold isotonic medium was added to the suspension and cells were again pelleted at 350 x g and the supernatant was removed. The pellet was resuspended in six mLs of Trizol LS (Ambion, Carlsbad, CA, USA) for RNA extraction. The remaining spleen half was immediately placed into Trizol LS for RNA extraction, and designated “**tissue**”. RNA extraction was performed according to the manufacturer’s instructions. RNAsin (40 units) was added to each RNA aliquot and stored in -80°C. RNA quality was assessed on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

### MiSeq Sequencing

One microgram of high-quality total RNA was used for RNA sequencing (RNA-seq) library construction using the TruSeq RNA Sample Preparation kit v2 (Illumina, San Diego, CA) with the following modification: one minute fragmentation time was applied to allow for longer RNA fragments. The obtained RNA-seq libraries were analyzed with the 2100 Bioanalyzer. The **tissue** sample described previously was then subjected to size selection and designated “**size selected**”. Selection of sequences 375-900 nucleotides (nt) in length (275-800 nt sequences plus 50 nt sequencing adaptors on each end of the cDNA) was performed using the Pippin Prep system



(Sage Science, Beverly, MA, USA). All libraries were then quantified with qPCR according to Illumina recommendations. The sequencing was performed at the Kansas State University Integrated Genomics Facility on the MiSeq personal sequencing system (Illumina) using the 600 cycles MiSeq reagent v3 kit (Illumina) according to Illumina instructions, resulting in 2 x 300 nt reads.

### **MiSeq Reference Mapping**

The bioinformatics workflow used in our study is outlined in Figure 1. FASTQ sequencing files were imported into CLC Genomics Workbench v9.5.1 (CLC bio, Aarhus, Denmark) (<https://www.qiagenbioinformatics.com/>). Data were cleaned in the CLC program to remove low quality and short sequences. Due to Illumina sequencing artifacts, the first 12 nt were removed from each sequence. Sequences were quality cleaned by retaining the longest region of the sequences with at least 97% of the sequence with Phred scores over 20. Sequences with fewer than 40 nt were removed. Reads remained with paired-end sequences, designated “**paired**”, and, in cases of overlapping sequence pairs, reads were merged, designated as “**merged**” (Figure 2.1A, light blue line). Sequences were merged using a match score of +1, mismatch cost of -2, a gap cost of -3, and a minimum score of 10. Cleaned paired (Figure 1a, purple lines) and merged (Figure 2.1A, red lines) sequences were mapped to specific C57BL/6 V gene sequences obtained from NCBI for the immunoglobulin heavy (IgH) and light (Igκ) chains. The paired and merged sequences were mapped using a match score of +1 and a mismatch score of -2. Additional putative antibody sequences were obtained by mapping sequences to the IgH and Igκ loci and the whole genome using the same scores. Mapped MiSeq sequences were combined and submitted to ImMunoGeneTics’s (IMGT) High-V Quest (Figure 2.1A, green lines) (54). Functionally

productive heavy chain sequences were imported into CLC and constant regions were determined using a motif search for the first 20 nt in each constant region that are provided in Table 2.1 (dashed line). Motifs were reassociated with their original sequences in Microsoft Excel for complete antibody (V[D]JC) identification. While two IgG subclass motifs were used to identify IgGs, they share partial sequence homology, therefore all IgG subclasses were combined resulting in an overarching IgG isotype. Kappa chain sequences were processed directly (Figure 2.1A, dotted line).

To collect a higher number of putative Ig sequences, our data handling workflow used multiple mapping processes which could result in the same sequence being submitted to IMGT multiple times. Failure to remove these duplicated sequences would lead to incorrect frequency assessments. Figure 2.2 outlines the procedure for duplicate sequencing read removal. Sequences were identified by their original MiSeq identification numbers for sorting. To retain the sequence with the most information and most accurate mapping, sequences were assessed based on functionality, constant region identification, V region score, and strand. Only one sequence per MiSeq identification number was retained and used for subsequent data compilation. Data from the remaining productive and unknown functionality antibody sequences were compiled for V-, D- (IgH only), and J-gene segment usage, CDR3 length, and CDR3 amino acid (AA) composition.

### **HiSeq Reference Mapping**

The MiSeq workflow described above was modified to analyze mouse liver transcriptomic data from the Rodent Research 1 (RR1) NASA validation flight provided by the NASA GeneLab program (<https://genelab-data.ndc.nasa.gov/genelab/>, Accession Numbers: GLDS-47, GLDS-48). These data include sequences from the livers of ground control and flight mice from two separate

cohorts, CASIS (GLDS-47) and NASA (GLDS-48), that were generated using Illumina HiSeq (1 x 50 nt reads). Raw sequencing reads were imported into CLC and quality cleaned as described with the exception of short read removal as quality cleaned reads were below the threshold utilized in the original workflow (Figure 2.1B). Reads were then mapped to the V $\kappa$  references identified above. Total V $\kappa$  mapping counts were collected and analyzed in Excel.

### **MiSeq and HiSeq Genome Mapping**

FASTQ files were imported into CLC and quality cleaned as described previously (Figure 2.1C). MiSeq reads were merged as described previously (Figure 2.1C, blue arrow). Paired and merged MiSeq and unpaired HiSeq reads (Figure 2.1C, purple arrow) were mapped using the RNA-Seq tool in CLC to the current mouse reference genome (GRCm38). A match score of +1, a mismatch score of -2, and insertion and deletion scores of -3 were used to map reads to the genome. Due to the short read lengths of the HiSeq data, V-gene segment usage was compiled directly after the RNA-Seq analysis (Figure 2.1C, green arrow). For MiSeq data, reads were collected, submitted to IMGT, duplicates removed and usage statistics compiled as described above (Figure 2.1C, dashed box).

### **NASA RNA Preparation and Sequencing**

Tissues from two sets of mice were analyzed from animals that were a part of the Rodent Research One (RR1) validation flight. The first set of spleens and livers were removed from five 35-week-old female C57BL/6Tac mice aboard the ISS 21-22 days after launch (CASIS Flight, SpaceX-4). Five 35-week old female mice housed in the ISS Environmental Simulator were processed similarly with a four-day delay (CASIS Ground Controls) (55). Spleens and livers were

placed in RNAlater (LifeTechnologies, Carlsbad, CA) for at least 24 hours at 4°C and then stored at -80°C while aboard the ISS, during transport, and upon return to Earth. The second set of tissues were isolated from seven 21-week-old female C57BL/6J mice that were euthanized aboard the ISS 37 days post-launch (NASA Flight, SpaceX-4). Carcasses were immediately frozen (-80°C) and after arriving on Earth, were dissected. Livers were preserved in RNAlater for at least 24 hours at 4°C and then frozen at -80°C. RNA was extracted from the tissues using Trizol (LifeTechnologies, Carlsbad, CA) according to the manufacturer's instructions. The resultant RNA was processed through an RNeasy mini column (QIAGEN, Hilden, Germany) as per manufacturer's instructions. RNA quality was assessed on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) and stored at -80°C. RNA isolated from liver tissue was sequenced on Illumina HiSeq with single reads of 50nt (1x 50nt). Additionally, Illumina MiSeq sequencing of splenic RNA isolated from three CASIS ground control animals was performed as described earlier.

## Results

### Size Selection Yields Highest Number of Antibody Sequences

Most studies of Ig gene-segment use frequency have used single-cell suspensions or sorted cells to isolate specific cell populations (5, 56-59). The advantage of these preparations is the exclusion of extraneous tissues and cells, which can enhance recovery of Ig sequences.

Our goal was to assess Ig gene segment usage in mice housed on the ISS. Due to limitations of animal and tissue handling on the ISS, only whole tissues would be available for analysis. To determine if we could obtain a sufficient number of Ig sequences from alternative sample preparations, we examined the total Ig sequence results from three different RNA preparation and sequencing treatment groups. The first two treatment groups comprised of RNA prepared from cells or whole tissue. A third treatment group was the same RNA from the tissue, which was subsequently size selected at 275-800 nt and sequenced independently. We selected this range of lengths because the IgH VDJ recombination sequences generally require 350-450 nt to gather information on V/D/J/C usage. Size selection also allows us to eliminate the inherent Illumina bias for short reads, while maintaining total transcriptome integrity for later data mining purposes.

Total sequence numbers generated were similar among the treatment groups with the cells treatment group resulting in 23.9 million reads, tissue treatment group with 25.9 million, and size selected treatment group with 21.5 million reads, as shown in Table 2.2. After cleaning, 18.7 million reads remained in the cells treatment group, 20.3 million in the tissue treatment group, and 12 million in the size selected group (Table 2.2). Mapping was performed as described in the materials and methods for both the IgH and Ig $\kappa$  loci and VH- and V $\kappa$ -gene segments. Locus mapping returned higher levels of probable Ig sequences than V-gene segment specific mapping. V-gene segment mapping of sequences to both locus and gene segment references from quality

cleaned reads was the lowest in the cells treatment group (1.56% IgH and 1.51% Igκ) and highest in the size selected treatment group (3.1% IgH and 2.69% Igκ).

After submission to IMGT and cleaning, antibody sequences were identified as potentially functional or of unknown functionality by IMGT. Unknown functionality sequences were comprised of partial sequences lacking CDR3 information to determine functionality. The total number of antibody sequences obtained for the IgH and Igκ was lowest in the cells treatment group and highest in the size-selected treatment group, (Table 2.2) mirroring the V-gene segment mapping results. Both productive and unknown IgH and Igκ antibody sequence counts were also lowest in the cells treatment group and highest in the size-selected treatment group. More unknown than productive sequences being identified in both IgH and Igκ. The size-selected treatment group produced both the highest number of productive antibody sequences and the highest total number of identified Ig sequences among treatment groups.

### **Comparison of Ig Gene Segment Usage Among Treatment Groups**

We compared IgH and Igκ gene segment frequency using multiple metrics across all three treatment groups; the first of which is V-gene segment usage. To assess the frequency of each VH- and Vκ- gene segment in normal mouse spleen, the total frequency of each V-gene segment was tabulated in Figures 2.3 and 2.4 as a percentage of the total repertoire for our cells, tissue, and size selected treatment groups. VH-gene segment usage was similar among all three treatment groups (Figure 2.3A). V-gene segment V1-80 was detected most frequently, followed by V1-18, and V1-26 gene segments. The gene segment V1-80 was ranked either first or third as a percentage of the total repertoire in all three treatment groups (Figure 2.3B). The gene segment V1-18 ranged between the first and third most frequently used gene segment. V1-26 was the second to fifth most

frequently used VH-gene segment. While gene frequency detection rankings among cells, tissue and size selected treatment groups were not identical, there was high similarity in overall VH-gene segment detection and in repertoire usage. Correlations between treatment groups produced an  $R^2$  of at least 0.7562 ( $p < 0.0001$ , data not shown) between the cells and size selected treatment groups. The  $R^2$  between cells and tissue treatment group was higher ( $R^2 = 0.8149$ ,  $p < 0.0001$ , data not shown). Tissue and size selected treatment groups had the highest correlation ( $R^2 = 0.9645$ ,  $p < 0.0001$  data not shown).

Kappa chain V-gene segment usage was also compared among the different treatment groups. Figure 2.4 shows the percent of repertoire for the top ten most abundant  $V_k$  gene-segments of each treatment group. There was significant overlap in the top  $V_k$  of each treatment group when assessed as either percent of repertoire detected (Figure 2.4A) or when ranked from highest to lowest frequency (Figure 2.4B). Greater similarities in  $V_k$  existed between tissue and tissue size selected treatments. Correlations between  $V_k$  in treatment groups produced an  $R^2$  of at least 0.8335 ( $p < 0.0001$ , data not shown), with tissue and size selected treatment groups having the highest correlation ( $R^2 = 0.9894$ ,  $p < 0.0001$ , data not shown).

The frequency of D- and JH-gene segment use in normal mice was also assessed. Figure 2.5 shows that the cells, tissue and size selected D- and JH-gene-segment usage frequency was similar. D1-1 was the most frequently discovered D-gene segment in all three treatment groups, comprising almost 30% of the repertoire (Figure 2.5A). Due to the short length of the D-gene segment, it was often difficult to properly determine which D-gene segment was used in an antibody. When a D-gene segment was identified, but was attributed to a non-strain-specific D gene, they were labeled “undetermined”. These D-gene segments were also very common, occurring between 26%-28% of the time, in all three treatment groups. Gene segments D2-4, D4-

1, D2-3, D2-5 and “no” D-gene segment (labeled NONE) were the next most frequent assignments, ranging from about six percent to eight percent of D-gene frequency.

We found that JH-gene segment usage was the same for the cells, tissue and size selected treatment groups (Figure 2.5B). JH2 was the most frequently used J-gene segment followed by JH1, JH4, and JH3, respectively. Gene segment usage in kappa chains was also similar in all three treatment groups, with J $\kappa$ 1 as the most frequently used, followed by J $\kappa$ 5, J $\kappa$ 2 and J $\kappa$ 4 respectively (Figure 2.5C). When less than six nucleotides from the J-gene segment were identified, they were marked as <6 nt.

Five heavy chains, IgA, IgD, IgE, IgG (all subfamilies), and IgM are part of the normal mouse Ig repertoire. Almost 80% of the repertoire used the IgM constant region (Figure 2.5D). IgD, IgA, and IgG were detected at frequencies between three percent and 12 percent of the total repertoire and were evenly distributed among cells, tissue and size selected treatment groups. IgE was not detected in any of the treatment groups.

### **CDR3 AA Sequence Determination**

CDR3 is highly variable and it may be critical in determining antigen specificity (6). Therefore, we assessed individual CDR3 frequency from each treatment group. The top five most common CDR3s from each treatment group for the heavy chain were compiled and shown in Figure 2.6A, resulting in a total of 10 unique CDR3s among treatment groups. The tissue and size selected treatment groups contained all of the most common CDR3s, however, the cells treatment group lacked the CARGIYYGSYFDYW sequence, which ranked as the second most common in the tissue data set (Figure 2.6B). We detected one hundred sixty-four CDR3 AA sequences in all three data sets at least once (Figure 2.6C). The tissue and size selected treatment data sets shared



607 CDR3 sequences. Thirty-eight and 82 CDR3 AA sequences were shared between cells and tissue and cells and size selected treatment groups, respectively. Each treatment group data set also contained a large number of unique CDR3 reads.

We found overlap among the top five kappa chain CDR3 of each treatment group (Figure 2.6D). All top CDR3 sequence were found among all three datasets and CDR3 sequences appeared in at least the top 78 CDR3 sequences of the other treatment groups out of 2814 unique CDR3 that were identified (Figure 2.6E). Figure 6F shows the diversity of kappa chain CDR3 sequences. The total number of CDR3 amino acid sequences that were unique to each treatment groups was 441, 811, and 848 in cells, tissue, and tissue size selected treatment groups, respectively. Three-hundred and eighty-one unique kappa chain CDR3 sequences are shared among all three treatment groups.

### **Application of MiSeq Workflow to RR1 HiSeq Data**

The MiSeq workflow was adapted to process the liver Illumina HiSeq data from the RR1 validation flight. Due to short read length (38 nt), only V-gene segment usage was assessed. Due to low IgH read numbers, only V $\kappa$  information is presented. The V $\kappa$  percent of repertoire from each HiSeq RR1 mouse cohort (CASIS Ground, CASIS Flight, NASA Ground, and NASA Flight) was compared to the V $\kappa$  percent abundance of the cell, tissue, and size selected HiSeq datasets. Table 3 shows poor correlation between reference mapped RR1 HiSeq cohorts and the MiSeq datasets described in the previous section. As all mice represented in this comparison are C57BL/6 mice, though ages and experimental conditions varied, we were concerned that the lack of concurrence in V $\kappa$ -gene segment usage of the RR1 mice and those used in the workflow discussed above may reflect the bioinformatic treatment of the data. To test this hypothesis, we modified the workflow to map sequencing reads to the entire *Mus musculus* genome rather than mapping reads

to V $\kappa$  reference sequences, as genome mapping is commonly employed in transcriptomics analysis. This bioinformatic treatment yielded a higher correlation with the MiSeq datasets. Table 3 shows that the distribution of V $\kappa$  percent abundance was dependent on the mapping technique used for the HiSeq datasets.

### **Reference-Locus Mapping Comparable to Whole Genome RNA-Seq Methods**

Because the data obtained from the HiSeq data set using reference mapping techniques were less correlative to V $\kappa$ -gene segment use to our previously obtained MiSeq data for normal mice we were concerned that our initial bioinformatics techniques for MiSeq data may not be appropriate. To validate the accuracy of our bioinformatic treatments of the sequencing data that were submitted to IMGT, two different mapping methods were compared using Illumina MiSeq data from CASIS ground control animals. The reference mapping approach, used previously, mapped sequences to both the IgH V-gene segments (251 segments) and the entire IgH locus (2.8Mb) obtained from NCBI (NC\_000078.6, 113258768 to 116009954) or Ig $\kappa$  V-gene segments (164 segments) and the entire Ig $\kappa$  locus (3.2Mb) obtained from NCBI (NC\_000072.6, 67555636 to 70726754). Therefore, we used the whole genome mapping outlined above to compare to our reference mapping. Results were obtained by using the RNA-Seq analysis tool in CLC to map reads to the entire genome with the IgH and Ig $\kappa$  loci selected for submission to IMGT. The IMGT output was processed similarly for both (genome *vs.* reference) mapping strategies. The median frequency of all VH- and V $\kappa$ -gene segments was compared if it was detected in at least one of the three animals and the data were compiled for both reference- and genome-mapping options. V-gene segments not detected in a treatment group were assigned a “zero” frequency. Assessment of the median frequencies of the two methods by linear regression in Figure 2.7 revealed that the

frequency data for VH-gene segment usage was very similar regardless of the mapping technique ( $R^2 = 0.9973$ ,  $p = <0.0001$ ) (Figure 2.7A). There was also a strong correlation of V $\kappa$  usage between the two methods ( $R^2=0.9991$ ,  $p<0.0001$ ) (Figure 2.7B). Comparisons of D-gene segment, J-gene segment, constant region frequency and CDR3 lengths were also highly correlated using both techniques (data not shown). Therefore, we are confident that our reference mapping bioinformatics strategy is providing an accurate picture of gene segment usage.

## Discussion

Spaceflight presents unique difficulties in the collection, preparation and preservation of cells and tissues. Normal preparation methods such as the creation of single-cell suspensions are difficult and normal tissue preservation methods such as the use of liquid nitrogen for flash-freezing are unavailable, leading to the preservation of tissue in RNAlater followed by long term storage at -80°C. In an effort to determine the acceptability of whole tissue preparations compared to more traditional single cell suspensions, we examined the differences in Ig sequences obtained from both treatment groups. We were concerned that tissue isolation methods may introduce artifacts into the data since many studies specifically focus on single cell suspensions, often sorted, to isolate B cells specifically (56, 58, 60, 61). While animals utilized in this study were not specifically challenged, the presence of plasma cells and plasmacytes, which produce several orders of magnitude more immunoglobulin transcripts, cannot be excluded and represents a weakness of our workflow for the analysis of naïve repertoires. Similarly, underrepresentation of subpopulations with limited stability may lead to other biases (62, 63).

Our data indicate that comparable results were obtained from both the tissue and the cells treatment groups as there were strong correlations in V-gene segment usage. Due to the combinatorial nature of CDR3, the level of shared sequence identity was encouraging among the three treatment groups. It should be noted, however, that even within the top H-CDR3, a higher degree of similarity was found in the tissue and size selected treatment groups than was found in the cells treatment group. While the high level of similarity between the tissue and size selected cohort is to be expected given that the size selected treatment was a subset of the tissue RNA, the deviation in the cells treatment group could possibly result from an unbalanced sampling of B-cell subsets depending on the portion of the spleen that was utilized during sample preparation, which

would be exacerbated by clonal expansion. For instance, marginal zone B cells have been shown to have a preference for short H-CDR3 amino acid motifs that lack n-nucleotide additions as compared to follicular B cells (64, 65).

In an effort to reduce Illumina bias towards short reads seen in the cells and tissue treatment groups, 10 months later, we sequenced the same tissue total RNA using size selection. The extended storage time after initial sequencing and additional freeze/thaw cycles are likely the cause of reduced numbers of post-cleaning reads due to RNA degradation. Nevertheless, the size-selected data set still provided the highest number of productive and unknown antibodies. Subsequent preparations have verified that size selection is helpful in providing the highest number of antibody sequences (data not shown). Therefore, we have chosen to include size selection in our protocol for NGS assessment of Ig-gene segment usage.

The antibody repertoires from numerous species have been analyzed using a variety of amplification, sequencing, and analysis techniques (60, 66). We chose to assess Ig gene usage without using amplification. Although many studies use amplification in order to obtain a higher number of reads, this may lead to bias into the repertoire (2, 66). Bias may be introduced due to primers or to errors created during the PCR reaction (5, 66). The large number of primers needed to amplify all the V genes in mice also presents some obstacles. Some have used 5' RACE with primers based on the constant region (66, 67). However, in order to amplify the entire repertoire, multiple 5' RACE primers are required, which still increases the chances for primer bias and increases costs significantly. Our goal was to examine the breadth of the antibody repertoire by gathering information about V, D, J, constant region usage and CDR3 composition. The detected V-gene segments and CDR3 sequences appear to parallel the repertoires reported using more

focused amplification methods. Therefore, we have a methodology for future studies that will examine the immune response to vaccination.

During the course of our studies, we had the opportunity to work with both HiSeq and MiSeq data. While Illumina sequencing (HiSeq and MiSeq) produces a higher volume of sequence reads, they are shorter and more prone to errors than sequencing with 454 or Sanger methods (66). However, Illumina sequencing has improved over time and is arguably now the NGS of choice. Our sequencing with Illumina MiSeq allowed us to obtain reads of up to 560 nt when forward and reverse sequencing ends were paired. This provides enough sequencing length to capture information from the V-gene segment to the constant region of both the heavy and light chains. We also found that as the sequences became longer, there was a drop off in sequencing quality, which has been previously reported (68).

Our workflow for Ig sequence isolation selected for sequences with the most information. This required the merging of overlapping read pairs to provide sequences long enough to identify the V, D, J, and constant regions. In order to collect the highest number of possible Ig sequences, we used multiple mapping techniques to both the V-gene segment and the locus in an effort to collect every possible Ig sequence. Preliminary workflow attempts found that each mapping technique isolated some unique sequences and that locus mapping resulted in a high number of “false positive” sequences. Subsequent sequence removal in Excel selected for antibodies containing the most data retaining productive antibodies with constant regions and high V-gene scores, a measure of the length and accuracy of match to the germline V gene segment. By utilizing multiple mapping methods and subsequent selection for the sequence with the most information, we are able to obtain a relatively large number of antibody sequences without introducing primer bias or PCR-induced sequencing errors.

To the best of our knowledge, this is the first data set of tissue based, non-amplified MiSeq analysis of the antibody repertoire. While our results are not a direct technique match to other published data sets, our normal mouse V-gene segment usage is consistent with the findings from other laboratories. For example, Collins used 5' RACE from the constant region followed by sequencing using 454 on a splenic cell suspension (57). Of the top ten VH-gene segments identified by Collins, we identified five of the same VH-gene segments within our top ten most frequently detected. All except the V1-59 gene segment were among the highest contributors to our repertoire (57). In addition, JH-gene segment frequencies were also relatively uniform with the J2-gene segment as the most frequently used (57). Yang performed sequencing on cell sorted B cells isolated from the spleen followed by amplification with primers specific for many, but not all, of the V heavy chains of mice and constant region primers to amplify V, D, J and part of the constant region (61). Their PCR products were then sequenced on the Illumina platform and aligned to known VH-gene segments (61). They identified V1-26 as their most common V-gene segment, which ranked between the second and fifth most common V gene in our data sets. V-gene segments V1-82, V1-64, and V1-55 were also identified as common V-gene segments in their analysis, all of which were frequently detected in our data (61). Kaplinski also examined sorted spleen cells, amplified with PCR and sequenced on MiSeq with 2 x 150 nt reads (58). Sequencing results were analyzed through idAB for identification (58). In contrast to our data analysis, Kaplinski examined only V gene segments found in productive antibodies, where we compiled all V gene segments identified in productive and unknown functionality sequences (58). Of the most common VH-gene segments provided, four, V1-80, V1-26, V1-53, and V1-82 were identified in our top ten grouping (58). Our data sets also isolated D1-1 as the most common D-gene segment. Kramer *et al.* examined sorted splenic follicular B cells, using IgM restricted PCR and sequenced using the

Sanger method (56). As we discovered, Kramer *et al.*'s most common VH family was V1, followed by V2 and V5 at relatively equal levels (56). In contrast, we found that the V6 gene-segment family was detected at a higher level than found in the Kramer analysis (56). The J4-gene segment was also used more than detected in our data set (56). We both identified D1 and D2 as the most common D gene-segment families.

We also compared our data to Igk gene family usage. Aoki-Ota assessed skewed V $\kappa$ -gene segment usage and V-J gene segment usage in unimmunized C57BL/6 mice using primer amplified total RNA of B cells from spleen, bone marrow and lymph node using 454 pyrosequencing (59). Their sequencing data was analyzed using the NCBI basic local alignment tool with reference sequences for V $\kappa$  and J $\kappa$  obtained from the IMGT data base. The top seven V-gene segments identified in their study were also found to be among the most abundant V $\kappa$  gene segments in all of our treatment groups. Additionally, V-J pairing of their top gene segments paralleled our data. Lu examined the effects of primer bias and mouse to mouse variation in V $\kappa$ - and J $\kappa$ -gene segments and CDR3 regions using primer amplified total RNA isolates of Balb/c splenic B cells on the 454 pyrosequencing platform (5). As with our study, sequencing reads were submitted to the IMGT HighV-quest tool, however, only functionally productive immunoglobulins were used in their analysis (5). In spite of the strain differences between our studies, of the V $\kappa$ -gene segments representing over one-percent of the antibody repertoire reported by Lu *et al.*, at least 80% of those gene segments appearing at a frequency 0.5% or higher in our analyses.

Although there was not 100% agreement among our study and the others, there was a high degree of consistency. Variations in data may result from sequencing and tissue isolation techniques and natural variation among animals, including mouse strain. In addition, since we did not amplify for V-gene segments, we likely may have missed more rare B-cell clones. Primer



biases in other studies may have also contributed to some of the differences. Nevertheless, it is clear that our approach provided an unbiased, representative sample of actively transcribing B cells.

Our group utilized liver RR1 sequencing data sequenced on the Illumina HiSeq platform (1 x 50 nt) that was available from the NASA Genelab project. The sequencing length was the largest limitation of these data. Our MiSeq data were sequenced in both directions at a length of 300 nt. Some paired-end reads also contained overlaps, allowing us to merge these sequences and provide reads up to around 560 nt. HiSeq sequencing reads were not of sufficient length to obtain CDR3 composition from the IMGT HighV-Quest tool, limiting the analyses that could be used to assess the antibody repertoire. Therefore, the applicability of publically available datasets to independent research questions is dependent upon the sequencing method used to acquire the data. For Ig gene repertoire studies, we recommend the use of sequencing methods that result in longer reads, though short reads may be useful for other research questions.

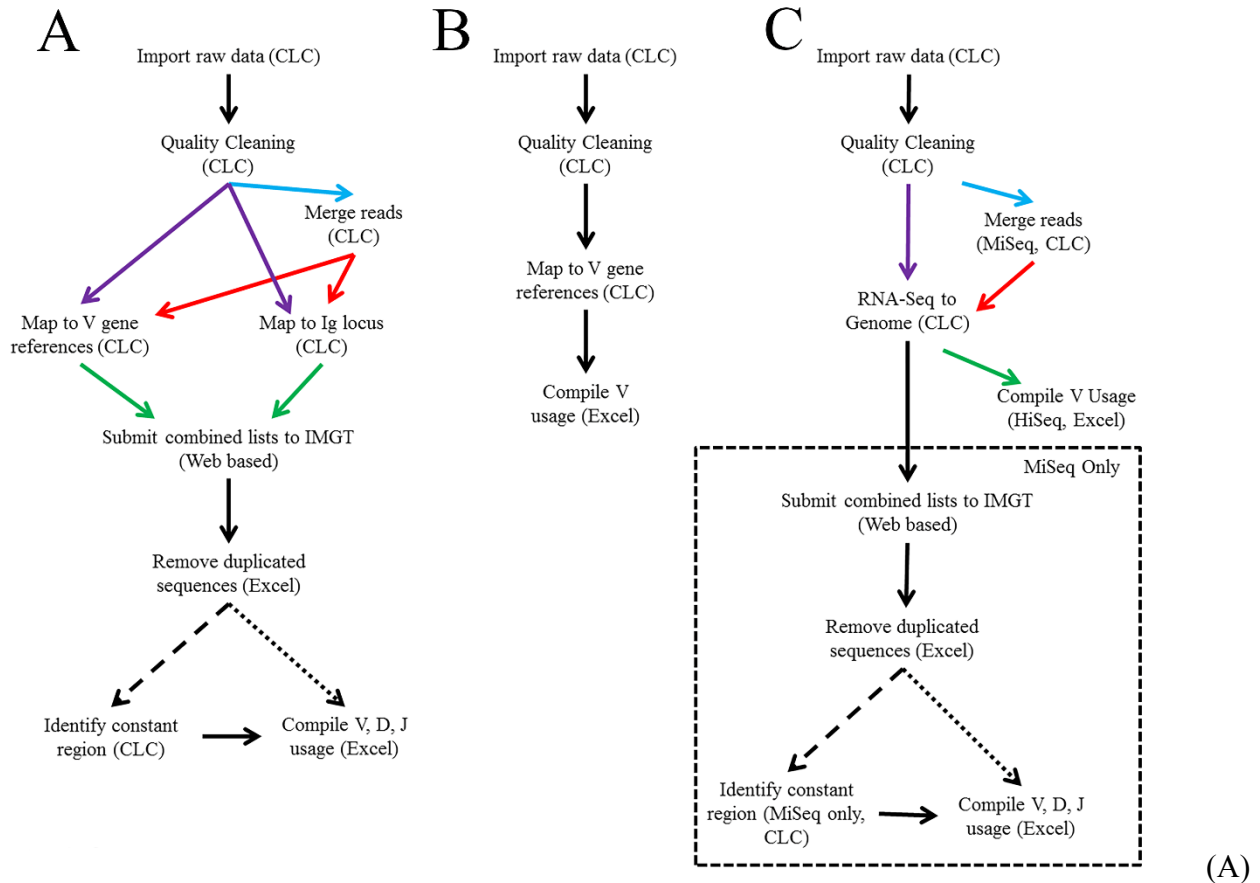
Initial comparisons to assess similarity of the different mouse cohorts showed a lack of correlation between the HiSeq RR1 data to the normal mouse MiSeq V $\kappa$  usage. We thought that part of the discrepancy may be from problems with the short HiSeq sequences, specifically when forced to align to V-gene segments when multiple matches are excluded. Mapping short HiSeq reads to the entire mouse genome remedied the inconsistencies observed between RR1 and normal mouse MiSeq data, likely due to the limitations of the RNA-Seq analysis employed. This demonstrates that the bioinformatic treatment of the data can impact results. We found that mapping longer MiSeq sequencing reads from RNA isolated from mice within the CASIS ground RR1 cohort to both the whole mouse genome and V $\kappa$  reference sequences yielded a strong correlation. This validates the applicability of the MiSeq workflow described in this work on

additional MiSeq datasets and reinforces that sequence read length must be taken into account when selecting bioinformatics methods.

In conclusion, our goals for this project were to examine the breadth of the antibody repertoire gathering information about V, D, J, and constant region usage and CDR3 composition and to lay the foundation for future studies that will examine the immune response to vaccination during space flight. We have determined that whole tissue preparations as will be available from the ISS will yield similar results when examining the antibody repertoire. We also determined that performing a size selection to isolate likely antibody sequences provided the highest number of Ig reads. A novel workflow using multiple mapping methods to characterize NGS data for Ig repertoire data was developed and genome and reference mapping methods were validated through the use of publically available datasets. This novel workflow can be used for future studies on the antibody repertoire regardless of whether they are ISS- or ground-based.

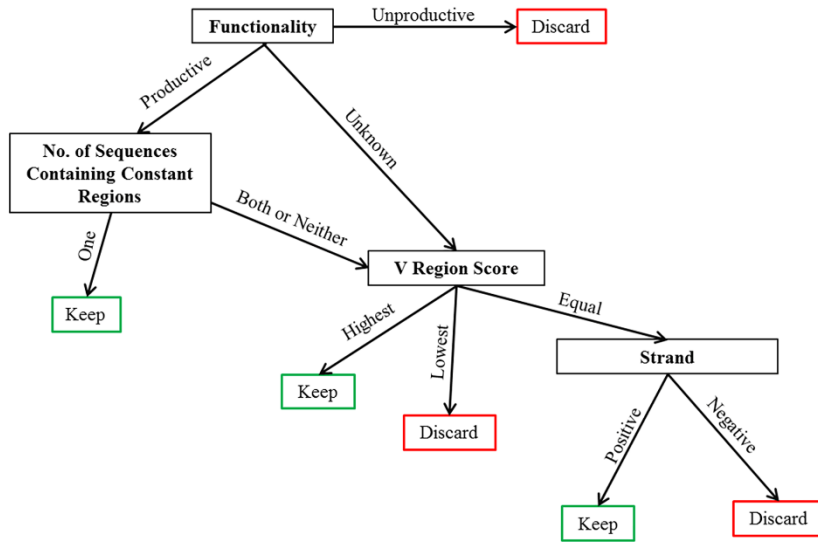
## Figures and Tables

Figure 2.1 Bioinformatic analysis workflows



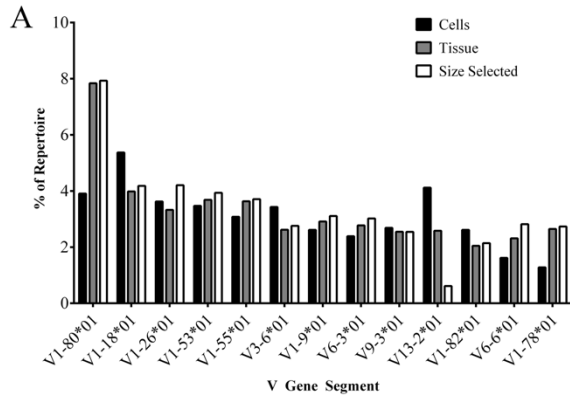
Workflow for MiSeq reference mapping strategy using CLC Genomics Workbench software, the ImmunoGeneTics (IMGT) data base and Excel. (B) Workflow for HiSeq reference mapping strategy. (C) Workflow for MiSeq and HiSeq referenced mapping strategy.

**Figure 2.2 Decision-making matrix to remove duplicate reads after IMGT processing**



Mapped sequences that were identified using Illumina sequence identification tags and sequences identified multiple times were removed as outlined.

**Figure 2.3 Top ten VH gene segments used among treatment groups**

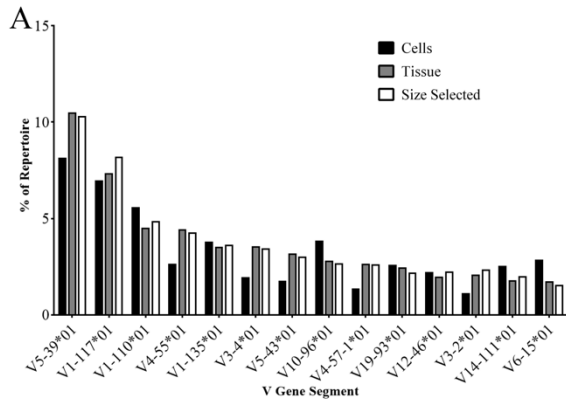


**B**

	Cells	Tissue	Size Selected
IGHV1-80*01	3	1	1
IGHV1-18*01	1	2	3
IGHV1-26*01	4	5	2
IGHV1-53*01	5	3	4
IGHV1-55*01	7	4	5
IGHV1-9*01	9	6	6
IGHV3-6*01	6	9	9
IGHV6-3*01	11	7	7
IGHV9-3*01	8	11	11
IGHV1-82*01	9	15	13
IGHV6-6*01	19	13	8
IGHV1-78*01	25	8	10
IGHV13-2*01	2	10	51

(A) The top ten VH gene segments for each treatment group are presented as a percent of repertoire with corresponding percent of repertoire in other treatment groups listed. (B) Top ten VH gene segments are listed by rank order (most frequent to least frequent). Dark red indicates higher rank moving to white, of lower rank. VH-gene segments with identical ranks are displayed as ties.

**Figure 2.4 Top ten V $\kappa$  used among treatment groups**

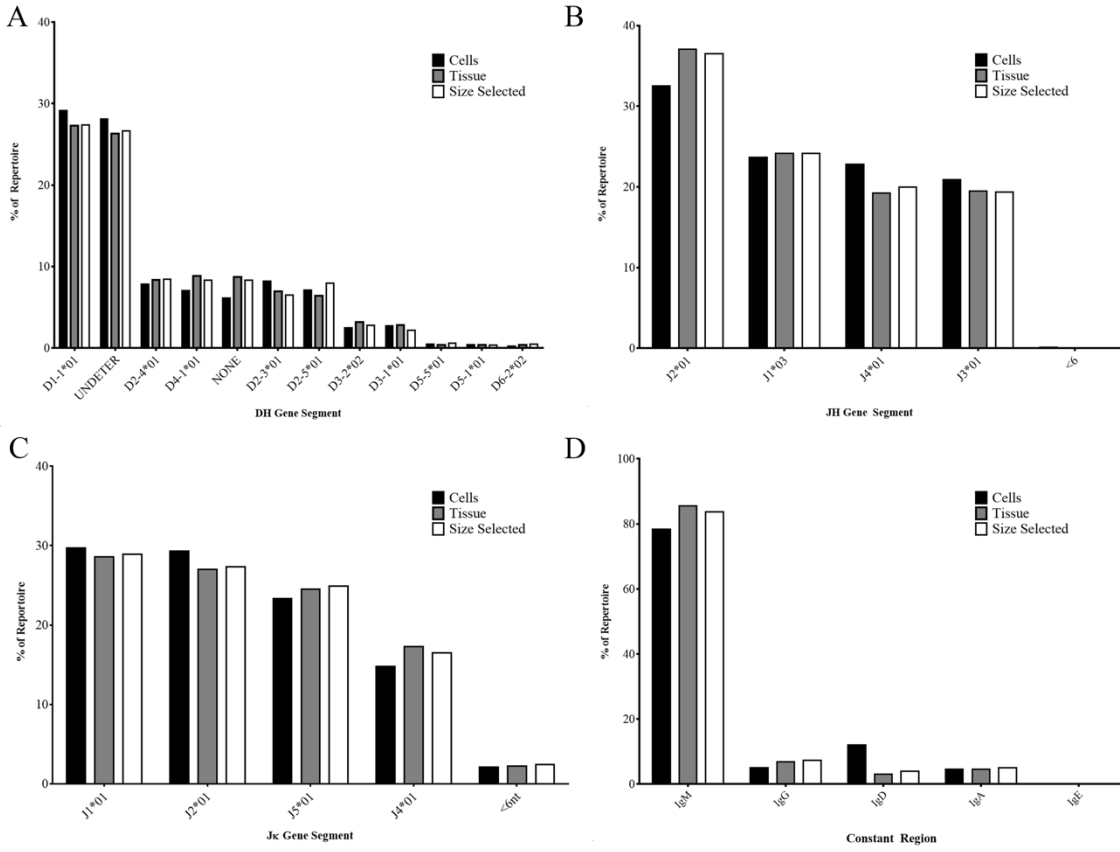


**B**

	Cells	Tissue	Size Selected
IGKV5-39*01	1	1	1
IGKV1-117*01	2	2	2
IGKV1-110*01	3	3	3
IGKV4-55*01	7	4	4
IGKV1-135*01	5	6	5
IGKV3-4*01	13	5	6
IGKV5-43*01	18	7	7
IGKV10-96*01	4	8	8
IGKV4-57-1*01	27	9	9
IGKV19-93*01	8	10	13
IGKV12-46*01	10	15	12
IGKV3-2*01	32	14	10
IGKV14-111*01	9	17	16
IGKV6-15*01	6	19	23

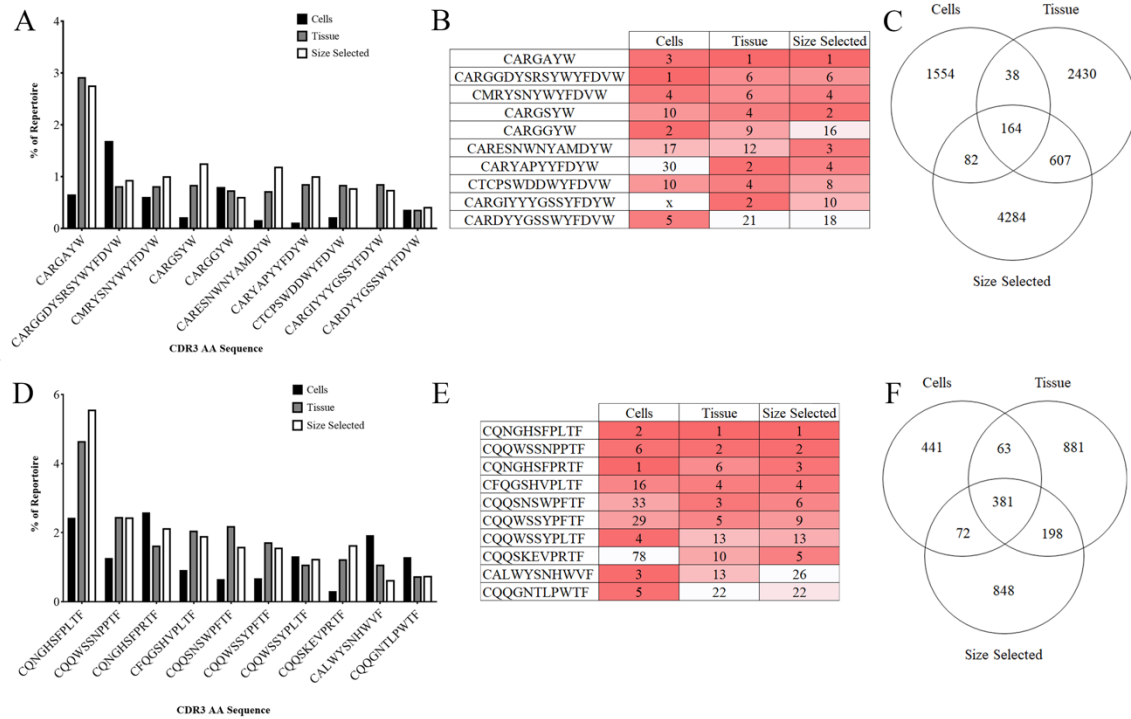
(A) The top ten V $\kappa$  gene segments for each treatment group are presented as a percent of repertoire with corresponding percent of repertoire in other treatment groups listed. (B) The top ten V $\kappa$  gene segments are listed by rank order (most frequent to least frequent). Dark red indicates higher rank moving to white, lower rank. VH-gene segments with identical ranks are displayed as ties.

**Figure 2.5 D, J, and heavy chain constant usage among treatment groups**



(A) DH gene segment usage by percent of repertoire. (B) JH gene segment usage by percent of repertoire. (C) Jk gene segment usage by percent of repertoire. (D) Heavy chain constant region usage by percent of repertoire.

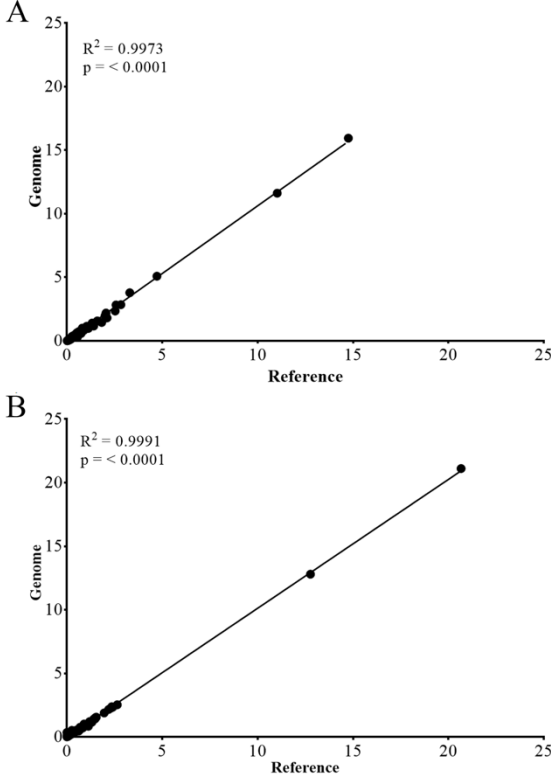
**Figure 2.6 CDR3 AA sequence usage among treatment groups**



(A) Top five most common heavy chain CDR3 AA sequence usage is presented as percent of repertoire. (B) Top five most common heavy chain CDR3 AA sequence usage is presented by rank. Dark red indicates higher rank moving to white, of lower rank. An x denotes that the AA sequence was not found. There was overlap in the top five IgH CDR3 among the treatment groups resulting in 10 most abundant CDR3 sequences. (C) Unique heavy chain CDR3 AA sequences identified within and among treatment groups. (D) Top five most common kappa chain CDR3 AA sequence usage is presented as percent of repertoire. (E) Top five most common heavy chain CDR3 AA sequence usage is presented by rank. There was overlap in the top five Igk CDR3 among the treatment groups resulting in 10 most abundant CDR3 sequences. (F) Unique heavy chain CDR3 AA sequences identified within and among treatment groups.



**Figure 2.7 Correlation of V gene segments between genome and reference mapping**



(A) Linear regression of median VH gene segment usage from genome and reference mappings.  $R^2=0.9973$ ,  $p<0.0001$ . (B) Linear Regression of median V $\kappa$  gene segment usage from genome and reference mapping.  $R^2=0.9991$ ,  $p<0.0001$ .

**Table 2.1 Sequences used for heavy chain identification**

<b>Constant Region</b>	<b>Motif Sequence</b>
IgA	GAGTCTGCGAGAAATCCCAC
IgD	GTAATGAAAAGGGACCTGAC
IgE	TCTATCAGGAACCCTCAGCT
IgG1/2b/2c	GCCAAAACAACAGCCCCATC
IgG3	AACAACAGCCCCATCGGTCT
IgM	TCAGTCCTTCCCAAATGTCT

Motifs used to determine the constant region of heavy chain Ig sequences.

**Table 2.2 Sequencing and mapping results from the cells, tissue, and size selected treatment groups**

	<b>Cells</b>	<b>Tissue</b>	<b>Size Selected</b>
Total Reads <sup>a</sup>	23.9 M	25.9 M	21.5 M
Post Cleaning <sup>a</sup>	18.7 M	20.3 M	12 M
IgH Mapped	278318	313194	327015
VH Mapped	12851	26559	42375
Igκ Mapped	261037	273562	264938
Vκ Mapped	20776	35719	64540
Heavy Chain Productive	2036	4991	8939
Heavy Chain Unknown	6139	11374	14047
Light Chain Productive	3439	6799	11595
Light Chain Unknown	6894	10462	12393

<sup>a</sup>M= Million sequencing reads

**Table 2.3 Comparison of mapping techniques in HiSeq datasets.**

		<b>Reference<sup>a</sup></b>	<b>Genome<sup>b</sup></b>	<b>Compared<sup>c</sup></b>
Cohort	N	R2 (p-value)	R2 (p-value)	R2 (p-value)
CASIS G	3	0.030 (.0637)	0.101 (0.0015)	0.011 (.027)
CASIS F	3	0.001 (.776)	0.216 (<.0001)	0.042 (.074)
NASA G	7	<0.001 (.854)	0.379 (<.0001)	0.013 (.262)
NASA F	7	0.004 (.521)	0.277 (<.0001)	0.006 (.476)

Mapping techniques were compared by assessing the correlation of V $\kappa$  usage between multiple HiSeq and Miseq datasets. HiSeq datasets included sequencing data from CASIS and NASA ground (G) or flight (F) RR1 mice. The comparison groups are as follows:

<sup>a</sup>V $\kappa$  gene segment usage from reference-mapped HiSeq data versus V $\kappa$  gene segment usage of MiSeq sample preparation datasets.

<sup>b</sup>V $\kappa$  gene segment usage from genome-mapped HiSeq data versus V $\kappa$  gene segment usage of MiSeq sample preparation datasets.

<sup>c</sup>V $\kappa$  gene segment usage from reference-mapped HiSeq Data versus V $\kappa$  gene segment usage of genome-mapped HiSeq data.

## References

1. de Bono, B., M. Madera, and C. Chothia. 2004. VH gene segments in the mouse and human genomes. *J. Mol. Biol.* 342: 131-143.
2. Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnol.* 32: 158-168.
3. Saul, F. A., and R. J. Poljak. 1992. Crystal structure of human immunoglobulin fragment Fab New refined at 2.0 Å resolution. *Proteins* 14: 363-371.
4. Haidar, J. N., W. Zhu, J. Lypowy, B. G. Pierce, A. Bari, K. Persaud, X. Luna, M. Snavely, D. Ludwig, and Z. Weng. 2014. Backbone flexibility of CDR3 and immune recognition of antigens. *J. Mol. Biol.* 426: 1583-1599.
5. Lu, J., T. Panavas, K. Thys, J. Aerssens, M. Naso, J. Fisher, M. Ryczyn, and R. W. Sweet. 2014. IgG variable region and VH CDR3 diversity in unimmunized mice analyzed by massively parallel sequencing. *Mol. Immunol.* 57: 274-283.
6. Xu, J. L., and M. M. Davis. 2000. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* 13: 37-45.
7. Durnova, G. N., A. S. Kaplansky, and V. V. Portugalov. 1976. Effect of a 22-day space flight on the lymphoid organs of rats. *Aviat. Space. Environ. Med.* 47: 588-591.
8. Grindeland, R. E., I. A. Popova, M. Vasques, and S. B. Arnaud. 1990. COSMOS 1887 mission overview: effects of microgravity on rat body and adrenal weights and plasma constituents. *FASEB J.* 4: 105-109.
9. Jahns, G., J. Meylor, T. Fast, N. Hawes, and G. Zarow. 1992. Rodent Growth, Behavior, and Physiology Resulting from Flight on the Space Life Sciences-1 Mission. *43rd Intl*

- Congress Astronautical Fed.* International Astronautical Federation, Washington, DC. 1-8.
10. Pecaut, M. J., S. J. Simske, and M. Fleshner. 2000. Spaceflight induces changes in splenocyte subpopulations: effectiveness of ground-based models. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 279: R2072-R2078.
  11. Wronski, T. J., M. Li, Y. Shen, S. C. Miller, B. M. Bowman, P. Kostenuik, and B. P. Halloran. 1998. Lack of effect of spaceflight on bone mass and bone formation in group-housed rats. *J. Appl. Physiol.* 85: 279-285.
  12. Allebban, Z., L. A. Gibson, R. D. Lange, T. L. Jago, K. M. Strickland, D. L. Johnson, and A. T. Ichiki. 1996. Effects of spaceflight on rat erythroid parameters. *J. Appl. Physiol.* 81: 117-122.
  13. Udden, M. M., T. B. Driscoll, L. A. Gibson, C. S. Patton, M. H. Pickett, J. B. Jones, R. Nachtman, Z. Allebban, A. T. Ichiki, R. D. Lange, and C. P. Alfrey. 1995. Blood volume and erythropoiesis in the rat during spaceflight. *Aviat. Space Environ. Med.* 66: 557-561.
  14. Chapes, S. K., S. J. Simske, G. Sonnenfeld, E. S. Miller, and R. J. Zimmerman. 1999. Effects of space flight and PEG-IL-2 on rat physiological and immunological responses. *J. Appl. Physiol.* 86: 2065-2076.
  15. Congdon, C. C., Z. Allebban, L. A. Gibson, A. Kaplansky, K. M. Strickland, T. L. Jago, D. L. Johnson, R. D. Lange, and A. T. Ichiki. 1996. Lymphatic tissue changes in rats flown on Spacelab Life Sciences-2. *J. Appl. Physiol.* 81: 172-177.
  16. Serova, L. V. 1980. Weightlessness effects on resistance and reactivity of animals. *Physiologist* 23: S22-S26.

17. Grove, D. S., S. A. Pishak, and A. M. Mastro. 1995. The effect of a 10-day space flight on the function, phenotype, and adhesion molecule expression of splenocytes and lymph node lymphocytes. *Exp. Cell. Res.* 219: 102-109.
18. Nash, P. V., and A. M. Mastro. 1992. Variable lymphocyte responses in rats after space flight. *Exp. Cell. Res.* 202: 125-131.
19. Lesnyak, A. T., G. Sonnenfeld, M. P. Rykova, D. O. Meshkov, A. Mastro, and I. Konstantinova. 1993. Immune changes in test animals during spaceflight. *J. Leukoc. Biol.* 54: 214-226.
20. Merrill, A. H., E. Wang, R. E. Mullins, R. E. Grindeland, and I. A. Popova. 1992. Analyses of plasma for metabolic and hormonal changes in rats flown aboard COSMOS 2044. *J. Appl. Physiol.* 73: 132S-135S.
21. Meehan, R., P. Whitson, and C. Sams. 1993. The role of psychoneuroendocrine factors on spaceflight-induced immunological alterations. *J. Leukoc. Biol.* 54: 236-244.
22. Stein, T. P., and M. D. Schluter. 1994. Excretion of IL-6 by astronauts during spaceflight. *Am. J. Physiol.* 266: E448-452.
23. Blanc, S., L. Somody, A. Gharib, G. Gauquelin, C. Gharib, and N. Sarda. 1998. Counteraction of spaceflight-induced changes in the rat central serotonergic system by adrenalectomy and corticosteroid replacement. *Neurochem. Int.* 33: 375-382.
24. Stowe, R. P., S. K. Mehta, A. A. Ferrando, D. L. Feedback, and D. L. Pierson. 2001. Immune responses and latent herpesvirus reactivation in spaceflight. *Aviat. Space Environ. Med.* 72: 884-891.
25. Stowe, R. P., D. L. Pierson, and A. D. Barrett. 2001. Elevated stress hormone levels relate to Epstein-Barr virus reactivation in astronauts. *Psychosom. Med.* 63: 891-895.

26. Stowe, R. P., C. F. Sams, S. K. Mehta, I. Kaur, M. L. Jones, D. L. Feeback, and D. L. Pierson. 1999. Leukocyte subsets and neutrophil function after short-term spaceflight. *J. Leukoc. Biol.* 65: 179-186.
27. Nash, P. V., I. V. Konstantinova, B. B. Fuchs, A. L. Rakhmilevich, A. T. Lesnyak, and A. M. Mastro. 1992. Effect of spaceflight on lymphocyte proliferation and interleukin-2 production. *J. Appl. Physiol.* 73: 186S-190S.
28. Lesnyak, A., G. Sonnenfeld, L. Avery, I. Konstantinova, M. Rykova, D. Meshkov, and T. Orlova. 1996. Effect of SLS-2 spaceflight on immunologic parameters of rats. *J. Appl. Physiol.* 81: 178-182.
29. Mandel, A. D., and E. Balish. 1977. Effect of space flight on cell-mediated immunity. *Aviat. Space Environ. Med.* 48: 1051-1057.
30. Sonnenfeld, G., M. Foster, D. Morton, F. Bailliard, N. A. Fowler, A. M. Hakenewerth, R. Bates, and E. S. Miller, Jr. 1998. Spaceflight and development of immune responses. *J. Appl. Physiol.* 85: 1429-1433.
31. Grigoriev, A. I., S. A. Bugrov, V. V. Bogomolov, A. D. Egorov, V. V. Polyakov, I. K. Tarasov, and E. B. Shulzhenko. 1993. Main medical results of extended flights on Space Station Mir in 1986-1990. *Acta Astronautica* 29: 581-585.
32. Taylor, G. R., L. S. Neale, and J. R. Dardano. 1986. Immunological analyses of U.S. Space Shuttle crewmembers. *Aviat. Space Environ. Med.* 57: 213-217.
33. Taylor, G. R., and J. R. Dardano. 1983. Human cellular immune responsiveness following space flight. *Aviat. Space Environ. Med.* 54: S55-59.



34. Cogoli, A., B. Bechler, O. Mueller, and E. Hunzinger. 1990. Effect of microgravity on lymphocyte activation. In *BioRack on Spacelab D1*. European Space Agency, Paris. 89-100.
35. Konstantinova, I. V., Y. N. Antropova, V. I. Legen'kov, and V. D. Zazhirey. 1973. Study of reactivity of blood lymphoid cells in crew members of the Soyuz-6, Soyuz-7 and Soyuz-8 spaceships before and after flight. *Space Biol. Aerospace Med.* 7: 48-55.
36. Hughes-Fulford, M. 1991. Altered cell function in microgravity. *Exp. Gerontol.* 26: 247-256.
37. Fuchs, B. B., and A. E. Medvedev. 1993. Countermeasures for ameliorating in-flight immune dysfunction. *J. Leukoc. Biol.* 54: 245-252.
38. Miller, E. S., D. A. Koebel, and G. Sonnenfeld. 1995. Influence of spaceflight on the production of interleukin-3 and interleukin-6 by rat spleen and thymus cells. *J. Appl. Physiol.* 78: 810-813.
39. Sonnenfeld, G., A. D. Mandel, I. V. Konstantinova, G. R. Tylor, W. D. Berry, S. R. Wellhausen, A. T. Lesnyak, and B. B. Fuchs. 1990. Effects of spaceflight on levels and activity of immune cells. *Aviat. Space Environ. Med.* 61: 648-653.
40. Gould, C. L., M. Lyte, J. Williams, A. D. Mandel, and G. Sonnenfeld. 1987. Inhibited interferon-g but normal interleukin-3 production from rats flown on the space shuttle. *Aviat. Space Environ. Med.* 58: 983-986.
41. Sonnenfeld, G., A. D. Mandel, I. V. Konstantinova, W. D. Berry, G. R. Taylor, A. T. Lesnyak, B. B. Fuchs, and A. L. Rakhmilevich. 1992. Spaceflight alters immune cell function and distribution. *J. Appl. Physiol.* 73: 191S-195S.

42. Bikle, D. D., J. Harris, B. P. Halloran, and E. R. Morey-Holton. 1994. Altered skeletal pattern of gene expression in response to spaceflight and hindlimb elevation. *Am. J. Physiol.* 267: E822-E827.
43. Allebban, Z., A. T. Ichiki, L. A. Gibson, J. B. Jones, C. C. Congdon, and R. D. Lange. 1994. Effects of spaceflight on the number of rat peripheral blood leukocytes and lymphocyte subsets. *J. Leukoc. Biol.* 55: 209-213.
44. Ichiki, A. T., L. A. Gibson, T. L. Jago, K. M. Strickland, D. L. Johnson, R. D. Lange, and Z. Allebban. 1996. Effects of spaceflight on rat peripheral blood leukocytes and bone marrow progenitor cells. *J. Leukoc. Biol.* 60: 37-43.
45. Meehan, R. T., L. S. Neale, E. T. Kraus, C. A. Stuart, M. L. Smith, N. M. Cintron, and C. F. Sams. 1992. Alteration in human mononuclear leucocytes following space flight. *Immunology* 76: 491-497.
46. Berry, C. A. 1970. Summary of medical experience in the Apollo 7 through 11 manned spaceflights. *Aerospace Med.* 41: 500-519.
47. Crucian, B., R. P. Stowe, S. Mehta, H. Quiariarte, D. Pierson, and C. Sams. 2015. Alterations in adaptive immunity persist during long-duration spaceflight. *Microgravity* 1: 15013.
48. Sonnenfeld, G. 2005. The immune system in space, including Earth-based benefits of space-based research. *Curr. Pharm. Biotechnol.* 6: 343-349.
49. Gridley, D. S., J. M. Slater, X. Luo-Owen, A. Rizvi, S. K. Chapes, L. S. Stodieck, V. L. Ferguson, and M. J. Pecaut. 2009. Spaceflight effects on T lymphocyte distribution, function and gene expression. *J. Appl. Physiol.* 106: 194-202.

50. Gaignier, F., V. Schenten, M. De Carvalho Bittencourt, G. Gauquelin-Koch, J.-P. Frippiat, and C. Legrand-Frossi. 2014. Three Weeks of Murine Hindlimb Unloading Induces Shifts from B to T and from Th to Tc Splenic Lymphocytes in Absence of Stress and Differentially Reduces Cell-Specific Mitogenic Responses. *PLoS one* 9: e92664.
51. Lescale, C., V. Schenten, D. Djeghloul, M. Bennabi, F. Gaignier, K. Vandamme, C. Strazielle, I. Kuzniak, H. Petite, C. Dosquet, J. P. Frippiat, and M. Goodhardt. 2015. Hind limb unloading, a model of spaceflight conditions, leads to decreased B lymphopoiesis similar to aging. *FASEB J.* 29: 455-463.
52. Fitzgerald, W., S. Chen, C. Walz, J. Zimmerberg, L. Margolis, and J. C. Grivel. 2009. Immune suppression of human lymphoid tissues and cells in rotating suspension culture and onboard the International Space Station. *In Vitro Cell. Dev. Biol. Anim.*
53. Crucian, B. E., S. R. Zwart, S. Mehta, P. Uchakin, H. D. Quiriarte, D. Pierson, C. F. Sams, and S. M. Smith. 2014. Plasma cytokine concentrations indicate that in vivo hormonal regulation of immunity is altered during long-duration spaceflight. *J. Interferon Cytokine Res.* 34: 778-786.
54. Alamyar, E., P. Duroux, M. P. Lefranc, and V. Giudicelli. 2012. IMGT((R)) 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.* 882: 569-604.
55. Figliozzi, G. M. 2014. NASA's New Rodent Residence Elevates Research To Greater Heights. [http://www.nasa.gov/mission\\_pages/station/research/news/rodent\\_research](http://www.nasa.gov/mission_pages/station/research/news/rodent_research).

56. Kramer, J. M., N. E. Holodick, T. C. Vizconde, I. Raman, M. Yan, Q. Z. Li, D. P. Gaile, and T. L. Rothstein. 2016. Analysis of IgM antibody production and repertoire in a mouse model of Sjogren's syndrome. *J. Leukoc. Biol.* 99: 321-331.
57. Collins, A. M., Y. Wang, K. M. Roskin, C. P. Marquis, and K. J. Jackson. 2015. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Phil. Trans. R. Soc. Lon. B* 370.
58. Kaplinsky, J., A. Li, A. Sun, M. Coffre, S. B. Koralov, and R. Arnaout. 2014. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc. Natl. Acad. Sci.* 111: E2622-2629.
59. Aoki-Ota, M., A. Torkamani, T. Ota, N. Schork, and D. Nemazee. 2012. Skewed primary Igkappa repertoire and V-J joining in C57BL/6 mice: implications for recombination accessibility and receptor editing. *J. Immunol.* 188: 2305-2315.
60. Greiff, V., U. Menzel, U. Haessler, S. C. Cook, S. Friedensohn, T. A. Khan, M. Pogson, I. Hellmann, and S. T. Reddy. 2014. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.* 15: 40.
61. Yang, Y., C. Wang, Q. Yang, A. B. Kantor, H. Chu, E. E. Ghosn, G. Qin, S. K. Mazmanian, J. Han, and L. A. Herzenberg. 2015. Distinct mechanisms define murine B cell lineage immunoglobulin heavy chain (IgH) repertoires. *eLife* 4: e09083.
62. Allman, D. M., S. E. Ferguson, V. M. Lentz, and M. P. Cancro. 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.

63. Allman, D., R. C. Lindsley, W. DeMuth, K. Rudd, S. A. Shinton, and R. R. Hardy. 2001. Resolution of three nonproliferative immature splenic B cell subsets reveals multiple selection points during peripheral B cell maturation. *J. Immunol.* 167: 6834-6840.
64. Carey, J. B., C. S. Moffatt-Blue, L. C. Watson, A. L. Gavin, and A. J. Feeney. 2008. Repertoire-based selection into the marginal zone compartment during B cell development. *J. Exp. Med.* 205: 2043-2052.
65. Schelonka, R. L., J. Tanner, Y. Zhuang, G. L. Gartland, M. Zemlin, and H. W. Schroeder, Jr. 2007. Categorical selection of the antibody repertoire in splenic B cells. *Eur. J. Immunol.* 37: 1010-1021.
66. Benichou, J., R. Ben-Hamo, Y. Louzoun, and S. Efroni. 2012. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135: 183-191.
67. Wang, Y., W. Chen, X. Li, and B. Cheng. 2006. Degenerated primer design to amplify the heavy chain variable region from immunoglobulin cDNA. *BMC Bioinformatics* 7 Suppl 4: S9.
68. Minoche, A. E., J. C. Dohm, and H. Himmelbauer. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* 12: R112.

# **Chapter 3 - Characterization of the Naïve Murine Antibody Repertoire Using Unamplified High-throughput Sequencing**

## **Abstract**

Antibody specificity and diversity are generated through the enzymatic splicing of genomic gene segments within each B cell. Antibodies are heterodimers of heavy- and light-chains which are encoded on separate loci. We studied the antibody repertoire from pooled, splenic tissue of unimmunized, adult female C57BL/6J mice, using high-throughput sequencing (HTS) without amplification of antibody transcripts. We recovered over 90,000 heavy-chain and over 120,000 light-chain immunoglobulin sequences. Individual V-, D-, and J-gene segment usage was uniform among the three mouse pools, particularly in highly abundant gene segments, with low frequency V-gene segments not being detected in all pools. Despite the similar usage of individual gene segments, the repertoire of individual B-cell CDR3 amino acid sequences in each mouse pool was highly varied, affirming the combinatorial diversity in the B-cell pool that has been previously demonstrated. There also appeared to be some skewing in the V-gene segments that were detected depending on chromosomal location. This study presents a unique, non-primer biased glimpse of the conventionally housed, unimmunized antibody repertoire of the C57BL6/J mouse.

## Introduction

B cells are an important part of the adaptive immune system, arising from hematopoietic stem cell precursors. These cells express surface immunoglobulin (Ig) receptors and secrete these same proteins as antibodies into the serum after differentiation into plasma cells (1, 2).

As B cells develop, they rearrange Variable- (V), Diversity- (D), and Joining- (J) gene segments, which combine with a constant region to form the antibody structure (3, 4). Antibodies consist of heterodimers of heavy- and light-chains (4). The heavy-chain is formed from V-, D-, and J-gene segments combined with a constant region (5), while light-chains lack a D-gene segment. (3, 6).

There are three complementarity determining regions (CDR). CDR1 and CDR2 are encoded in the V-gene segment. CDR3 consists of a combination of V-, (D-, heavy-chain), and J-gene segments (7). Of the CDRs, CDR3 contributes the most to binding specificity. Antibodies are further characterized by the constant region, or isotype, which is influenced by the stage of B-cell development and antigen specificity (8).

The total collection of antibody specificities present within an individual is known as the antibody repertoire. Diversity of the antibody repertoire results from four main components: the initial germ line (inherited), diversity from recombination of that germline, the imprecisions during V(D)J recombination, and somatic mutations (9-11). A snapshot of the antibody repertoire has been examined in many studies by high-throughput sequencing (HTS), and the antibody repertoire has been fully mapped in the zebrafish, due to their small size (12).

Repertoires can serve as a fingerprint or snapshot of the current immune-system status and these types of data have been used to explore the development of host defense to infectious disease (13, 14) (15-18), cancer (19-22), autoimmune disease (23, 24), and early disease detection (25).

With the development of HTS we are now able to detect the differences between or among B-cell repertoires such as B2 (adaptive antibodies) and B1 (natural antibodies) B cells (11) or memory and naïve repertoires (26, 27). HTS has accelerated the characterization of the widely differing human Ig haplotypes (28-32), and strain-specific gene segment usage in mice (33).

We are interested in the repertoire of B cells in mice and how it changes in response to antigen challenge. More specifically, our lab is interested antibody repertoire dynamics within the context of spaceflight. Due to the cost of these experiments, creating datasets that can be mined by our lab or others is important. The antibody repertoire is traditionally assessed through the amplification of Ig sequences that have been isolated from sorted B cell populations (34). While these practices increase the likelihood of recovering rare Ig sequences and allow for the dissection of the antibody repertoire by B-cell populations, cell sorting may not be possible within the design of certain experiments. During the development of methodology to do assess Ig-gene usage by mice subjected to space flight, we performed multiple HTS runs to validate sample preparation, bioinformatic methodology, and reproducibility (35). We present the data on the splenic repertoire of conventionally housed, unimmunized, unchallenged, adult C57BL/6J mice.

## **Materials and Methods**

### **RNA Extraction and Sequencing**

Tissue extraction and sequencing were performed as described previously (35). Briefly, spleens were collected from three independent pools of four, specific pathogen-free, female, C57BL/6J mice nine-to-eleven weeks old. Animals used in pool one were C57BL/6J mice raised in the laboratory animal care and services vivarium at Kansas State University (breeder stock renewed less than 2 years previously). Mice were fed LabDiet 5001 (LabDiet) and had access to



water and food *ad libitum*. Mice were maintained on a 12/12 light/dark cycle. Mice for pools two and three were purchased from Jackson Laboratories and allowed to acclimate in the vivarium for 22-31 days prior to sacrifice. Animal procedures were approved by the Institutional Animal Care and Use Committee at Kansas State University. Spleen tissue was processed immediately for RNA extraction with Trizol LS according to the manufacturer's instructions. Pool one contained RNA from one-half of the spleen tissue while pools two and three contained RNA extracted from complete spleens. Equal concentrations of total splenic RNA (RIN>8) were pooled and mixed from the four mice, resulting in three final pools. At least one microgram of RNA from each pool was submitted for size selection (275-800 nucleotides) and sequencing on Illumina MiSeq at 2x300 nucleotides as described previously (35). To avoid potential primer bias and maintain a dataset that could be further mined, we did not amplify Ig sequences.

## **Bioinformatics**

Sequence selection, mapping, and final processing was performed as outlined previously (35). Briefly, sequences were imported into CLC Genomics Workbench v9.5.1 (<https://www.qiagenbioinformatics.com/>) and cleaned to obtain high quality reads. Potential antibody sequences were isolated and submitted to ImMunoGenTics's (IMGT) High-V Quest for identification (36). Productive and unknown functionality sequences were identified via IMGT and used for subsequent analyses. Productive antibody sequences were defined as in frame and did not contain a stop codon, however binding abilities were not assessed. Unknown sequences did not contain enough sequencing information to determine functionality. Gene segments were identified using IMGT's nomenclature. We have implemented two procedural changes to further define the repertoire that are different from the workflow in chapter two (35). In this chapter, when

calculating the percent abundance in the repertoire, we also include V-gene segments where one or two possible V-gene segments were detected. When one single V-gene segment was detected in a sequence, it was assigned a value of one. When two potential V-gene segments were detected each gene segment was assigned a value of 0.5. The totals were then tabulated as described in Rettig *et al* (35). Additionally, CDR3 sequences that did not fit the C-xx-W motif for IgH (heavy-chain) were classified as unknown functionality, unless a class-switched isotype was detected. CDR3 sequences for Igκ (kappa-chain) that did not fit the C-xx-F motif were classified as unknown functionality. Sequencing reads containing hyperlengthy ( $\geq 25$  amino acids)  $\kappa$ -CDR3 that fit the C-xx-F motif were also removed from analysis as we believed they were falsely identified.

Initial heavy-chain nucleotide alignments were created with MAFFT (37) using portions of the germline and CDR3 nucleotide sequences provided by IMGT. The light-chain was aligned using a multiple sequence alignment in CLC. Sequences were sorted by identity compared to the germline and the sequence order was then adjusted to group similarly-aligned sequences. Nucleotide sequences of identical length were then isolated from the full alignment and aligned with each other while retaining all previously-inserted gaps.

### **V(D)J Pairing Frequency**

Pairing frequency was assessed in productive sequencing reads from both IgH and Igκ datasets. All pairing of V-gene segments was only assessed from productive IgH and Igκ sequencing reads, referred to as V<sub>H</sub> and V<sub>κ</sub>, respectively. Sequences identifying more than one possible V-gene segment were excluded from this analysis. For both IgH and Igκ, J- and D-gene segments designated as undetermined (U) were either not reported by IMGT, contained less than six nucleotides, or multiple gene segments were assigned to a single sequencing read. Total counts

from VJ pairings for heavy- and light-chains were tabulated and Circos graphs were generated using Circos Online (38).

### **Statistical Analysis and Graphics**

Linear Regressions were performed by comparing the percent of repertoire of V-gene segment or V(D)J combinations from pools 1 vs 2, 2 vs 3, and 1 vs 3 using the linear regression analysis tool in GraphPad (Version 6.0). Chi-square analysis of V-gene segment usage was performed on raw sequencing read counts using R version 3.3.1 (<https://www.r-project.org/>). All productive VH- and V $\kappa$ -gene segments listed in the National Center for Biotechnology Information (NCBI) *Mus musculus* reference assembly GRCm38.p5 were analyzed and gene segments not found in our datasets were assigned a read count of zero (<https://www.ncbi.nlm.nih.gov/grc/mouse>). These analyses were performed on each mouse pool separately by comparing the observed raw read count values of V-gene segments to an expected theoretical number of reads which was based on the null hypothesis that all V-gene segments will have the same number of raw reads. This value was determined by dividing the total number of sequencing reads observed in a mouse pool by the number of possible gene segments. The analysis of gene segment usage by chromosomal location was performed by dividing gene segments into four quadrants based on nucleotide position. A Chi-square analysis was performed on each mouse pool by summing the raw read counts of all gene segments containing a 5' nucleotide position within each quadrant and comparing the observed total reads within a quadrant to an expected theoretical number of reads for each quadrant which was based on the null hypothesis that the number of raw reads is not statistically different between defined quadrants. The expected raw-read count for each quadrant was determined separately by dividing the number of total raw reads

for a quadrant by the number of total possible gene segments and multiplying by the number of genes in in that quadrant.

Percent of repertoire values were determined by dividing sequencing reads corresponding to each gene segment, constant region, or CDR3 length by the total number of gene segments, constant regions, or CDR identified in each mouse pool for normalized comparison between pools. Percent of repertoire for these variables were displayed as bar graphs, generated in GraphPad v6.0, or as heatmaps, generated in Microsoft Excel. In addition to visualization of gene segment combinations through Circos graphs, the percent of repertoire for gene segment pairings was also visualized using the bubble chart tool in Microsoft Excel.

## **Results**

### **VH- and V $\kappa$ -Gene Segment Usage**

We obtained between 8,714 and 11,200 IgH individual productive reads, and between 14,271 and 28,756 individual IgH reads of unknown functionality as identified by IMGT HighV-Quest (Table 3.1). Between 12,199 and 15,115 individual Ig $\kappa$  productive, and 13,264 and 36,741 individual Ig $\kappa$  reads of unknown functionality were identified by IMGT HighV-Quest (Table 3.1). Overall, we identified 147 VH- and 100 V $\kappa$ -gene segments within the repertoires of our mouse pools (Appendix A.2). As a general trend, the three pools resulted in similar frequencies and similar ranks for V-gene usage among groups. The ten most common V-gene segments from each pool were compiled, resulting in a total use of 14 VH-gene segments and 16 V $\kappa$ -gene segments (Figure 3.1).

The most common VH-gene segment in pools one and two was V1-80 (Figure 3.1A). V1-80 was the sixth most common gene segment used in pool three. V6-3 was the most common VH-

gene segment in pool three, but ranked seventh and sixth in pools one and two, respectively (Figure 3.1B). V1-26 was the next most common VH-gene segment, ranking second in pools one and two, and third in pool three. Among the top ten most common VH-gene segments, most gene segments ranked between first and 17th within their pools, however, three outliers were found within these groups. V1-50 ranked 23rd in pool one, but it was ninth and second in pools two and three, respectively. V1-78 was ranked 31st in pool two, but ranked tenth and 17th in pools one and three, respectively. V1-18 was ranked 31st in pool three but third and fourth in pools one and two, respectively. VH-gene usage between pools was well correlated (1 vs 2  $R^2=0.8481$ , 2 vs 3  $R^2=0.7054$ , 1 vs 3  $R^2=0.5842$ , all  $p<0.0001$ ).

Sixteen V $\kappa$ -gene segments were among the top 10 most abundant V $\kappa$  of the repertoire in at least one of the three mouse pools, with six V $\kappa$ -gene segments appearing in the top ten of all three mouse pools (Fig. 3.1). Pool one appeared enriched for V5-39, comprising 10.1% of the repertoire as compared to 1.3% in pool two and 3.5% in pool three (Figure 3.1C). Excluding this difference, V1-110, V1-117, and V4-55 were the three most abundant gene segments in all three mouse pools. The lowest ranking of the most abundant V $\kappa$  in any of the mouse pools was V5-39, ranking 29<sup>th</sup> in pool 2, while ranking first and sixth in pools one and three, respectively (Figure 3.1D). All top V $\kappa$  that showed any variation in abundance among pools were still within in the top thirty V $\kappa$ -gene segments. V $\kappa$  usage between pools was correlated, although not as highly as VH (1 vs 2  $R^2=0.4701$ , 2 vs 3  $R^2=0.6848$ , 1 vs 3  $R^2=0.6306$ , all  $p<0.0001$ ).

With 182 productive VH and 151 productive V $\kappa$  described in NCBI, each gene segment would be expected to appear as part of the repertoire roughly 0.55% and 0.66% of the time for VH and V $\kappa$ , respectively, if gene segment usage was random (Figure 3.2). When we assessed usage

frequency, there appeared to be a non-random distribution of both VH- (Figure 3.2A) and V $\kappa$ - (Figure 3.2B) gene segments (Chi-square analyses;  $p < 0.0001$ ).

To assess gene segment usage by chromosomal location VH and V $\kappa$  were evenly divided into four quadrants from 5' to 3' (Q1, Q2, Q3, Q4). Gene segments appearing at an average of greater than three percent, or approximately five times the expected percent of repertoire, were identified. Only one VH-gene segment in Q1 (V1-80) represented over three percent of the repertoire. Two gene segments in Q2 (V1-53 and V1-26) represented at least three percent of the VH-gene repertoire. Q3 had one gene segments (V6-3) that was over-represented. In contrast, no genes assigned to Q4 made up over 3% of the repertoire.

When we analyzed V $\kappa$ -gene segment usage in a similar fashion, no gene segments appeared as more than three percent of the repertoire in Q2. One gene segment appeared over three percent presence in Q4 (V3-4). Two gene segments had a greater than three percent presence in Q3 (V4-55, V5-39) whereas three gene segments made up more than three percent of the repertoire in Q1 (V1-135, V1-117, V1-110).

Chromosomal location in the IgH and Ig $\kappa$  was also assessed by nucleotide position rather than even distribution into four quadrants because gene-segment spacing was not evenly distributed. Based on 5' nucleotide position, Q1-4 contained 50, 50, 40 and 28 VH-gene segments or 33, 34, 44 and 40 V $\kappa$ -gene segments, respectively. Since V-gene segment usage appears skewed, we tested whether the total expression of V-gene segments within a quadrant defined by nucleotide position was proportional to the number of gene segments located within the quadrant by Chi-square analysis in all mouse pools for both IgH and Ig $\kappa$ . We found that the total percent expression within each quadrant was not proportional to the number of gene segments for both IgH and Ig $\kappa$ , suggesting that V-gene expression may be influenced by chromosomal location ( $P < 0.05$ , all pools).

## **DH-, JH-Gene Segment and IgH Constant Region Usage**

We identified ten different D-gene segments used in our repertoires (Figure 3.3A). We also added one additional category for our analyses, termed “undetermined”. This label was applied to D-gene segments that were assigned by IMGT to non-C57BL/6J genes and antibody sequences containing a V- and J-gene segment, but not containing an identifiable D-gene segment. Due to the very short length of D-gene segments combined with alterations during recombination, D-gene segments were bioinformatically difficult to identify.

For all three groups, the D1-1 gene segment was the most common segment identified comprising 26-28% of the repertoire. Undetermined D-gene segments, however, made up a large part of the D-gene segment repertoire, comprising 32-35% of the repertoire. D2-3, D2-4, D4-1, and D2-5 were found in similar frequencies ranging from six to fifteen percent of the data set. D3-2, D3-1, D6-2, D5-1, and D5-5 were found at low levels in all data sets; comprising under three percent of the total repertoire.

Four JH-gene segments were identified, with JH2 being the most common among all three groups (Figure 3.3B). The remaining J-gene segments, JH4, JH3, and JH1, were found at similar levels among groups totaling between 19% and 27% of the repertoire.

IgM was overwhelmingly the most commonly identified constant region making up between 78% and 84% of the total repertoire (Figure 3.3C). IgG was the next most common between seven and eleven percent of the total repertoire. IgA and IgD were relatively rare totaling between two and six percent of the repertoire. IgE was only detected in pool three at less than one percent. It was not detected in pools one and two.

## **J $\kappa$ -Gene Segment Usage**

A total of four J $\kappa$ -gene segments were identified, with similar distribution of J $\kappa$ -gene segments between mouse pools (Figure 3.3D). Due to the even distribution of the three most abundant J $\kappa$ , the ranking of each gene segment only varied slightly among the three mouse pools. Within each mouse pool, there was a small portion of J $\kappa$  that contained too few nucleotides to be assigned to a specific gene segment (1.7-2.4%).

## **IgH- and Ig $\kappa$ - Gene Segment Combinations**

VH, DH, and JH family combination frequency was examined. Some preferential bias for specific gene segments seemed to exist (Figure 3.4A). For example, the JH4/DH2 combination appeared at a high frequency with VH1 (4.5% of repertoire), but not with any other VH gene family to the same degree. IgH gene segment recombination frequency correlated with gene segment abundance. VH1, which contains over half of all possible V-gene segments, also was the most commonly used VH family, which is seen as the dominant band in the Circos plot (Figure 3.4B).

The pairing of V $\kappa$  families to individual J $\kappa$  was also assessed (Figure 3.4C-D). Overall, the pairing of V $\kappa$  families with J $\kappa$  appeared random, however, certain V $\kappa$  families preferentially paired with specific J $\kappa$ -gene segments. For example, V4 paired less efficiently with J1, while V3 paired more efficiently with J1 (Figure 3.4C). Unlike VH, no single V $\kappa$  family was exceedingly dominant. Although V4 was the most represented gene family, its expression level was close to that of the second next most prominent gene families, which varied by mouse pool as shown in the Circos plot (Figure 3.4D).



The percent of repertoire that each VJ-gene segment combination comprised within each mouse pool was compared by linear regression. Mouse pools showed modest correlation levels of VJ-gene segment recombination frequency in IgH (1 vs 2  $R^2=0.5547$ , 2 vs 3  $R^2=0.4386$ , 1 vs 3  $R^2=0.3933$ , all  $p<0.0001$ ) and Ig $\kappa$  (1 vs 2  $R^2=0.2319$ , 2 vs 3  $R^2=0.3674$ , 1 vs 3  $R^2=0.4543$ , all  $p<0.0001$ ), with some enrichment for certain combinations within each mouse pool. V-(D)-J combinations were displayed in bubble charts generate a visual comparison of pairing (Figure 3.4A and 3.4C).

### **IgH and Ig $\kappa$ CDR3**

The average IgH CDR3 (H-CRD3) length of all three data sets was 11 amino acid (AAs) long (Figure 3.5A). The lengths of the H-CDR3s followed a normal distribution except all three groups were enriched for five AA H-CDR3s. H-CDR3 AA length ranged from one to twenty-three amino acids in length with 11 AAs being the average for all three pools. Ig $\kappa$  chain CDR3 length was conserved at nine AAs, comprising 87-90% of the repertoire (Figure 3.5B). While nine AAs was the most frequent  $\kappa$ -CDR3 length, one  $\kappa$ -CDR3 with a length of seven AAs was observed in the top CDR3 sequences. The distribution of CDR3 lengths was relatively even among pools for both IgH and Ig $\kappa$  (Appendix A.3). Four hyperlengthy  $\kappa$ -CDR3 sequences of that still fit the conserved C-xx-F motif were identified within the mouse pools (data not shown), while no such hyperlengthy H-CDR3 sequences fitting the C-xx-W motif were identified.

A total of 17,216 unique H-CDR3 AA sequences were identified among all three data pools. Among those identified, the majority (16,783) were identified in only one pool (Figure 3.6A). Of the remaining CDR3s, 358 were identified in only two pools and 75 were identified in all three pools.

Interestingly, many of these H-CDR3s, though found in all three pools, were not necessarily common H-CDR3s. Only one H-CDR3, CARGAYW, was found among the top ten most common H-CDR3s of each pool. One additional H-CDR3, CARDYYGSSWYFDVW, was found in the top 10 of pools two and three. Of the 75 total H-CDR3s that appeared in all three pools, frequencies varied drastically, from being the most common to only being detected once (Figure 3.6A). Of the top five most common H-CDR3s in each data set, only three, CARGAYW, CARGGYW, and CMRYSNYWYFDVW occurred in all three data sets (Figure 3.6B). CARGSYW occurred in pools one and two, CARRWLHYAMDYW in pools two and three, and CARYAPYYFDYW in pools one and three. The remaining most common H-CDR3s occurred in only one pool. A heatmap of all 75 shared H-CDR3 is shown in Appendix A.4.

There were 2,876 total unique  $\kappa$ -CDR3 amino acid sequences identified among all three mouse pools (Figure 3.6C). While there were 2,088 individual  $\kappa$ -CDR3 amino acid sequences that were unique to the individual mouse pools, there were also 436  $\kappa$ -CDR3 shared among all three pools. None of these 436 shared  $\kappa$ -CDR3 were found within the top 10  $\kappa$ -CDR3s of all three mouse pools (Figure 3.6D). One  $\kappa$ -CDR3 was among the top 10 for mouse pools one and two (CQQWSSYPPTF). Two  $\kappa$ -CDR3 were among the top 10 for mouse pools two and three (CQQWSSYPLTF, CQQYNSYPLTF). Three  $\kappa$ -CDR3 were among the top 10 for mouse pools one and three (CQNGHSFPLTF, CQQSNEDPRTF, CQQWSSNPPTF). A heatmap of all 436 shared  $\kappa$ -CDR3 is shown in Appendix A.4.

There were 13 total  $\kappa$ -CDR3 sequences that were found within the top 5  $\kappa$ -CDR3 for each mouse pool (Figure 3.6D). There were no CDR3 sequences that were found within the top 5  $\kappa$ -CDR3 in all three mouse pools. One  $\kappa$ -CDR3 sequence was among the top 5 for mouse pools one and two (CQQWSSYPPTF). Two  $\kappa$ -CDR3 were among the top 5 for mouse pools two and three

(CQQWSSYPLTF, CQQYNSYPLTF). Two  $\kappa$ -CDR3 were among the top 5 for mouse pools one and three (CQNGHSFPLTF, CQQWSSNPPTF).

### **Comparison of Alignments of CDR3s**

To assess the heterogeneity in B-cell idiotypes created by the differential splicing of Ig genes, we compared B cells that used the same V-, D- and J-genes. Two gene combinations containing complete V-, D-, and J-segments common to top 35 gene combinations found in all three mouse pools were selected (Figure 3.7A-C, Appendix A.5).

One heavy-chain VDJ-gene combination displaying a CDR3 region of variable length was selected from the 15 most common gene combinations among the three mouse pools and aligned to its germline sequence. From the full alignment, one short (four to eight AAs, Figure 3.7A), one medium (11 AAs, Figure 3.7B), and one long (14 AAs, Figure 3.7C) selection of nucleotide sequences were isolated and compared. Although the three groups were encoded by the same V-, D-, and J-gene segments, gene segment representation across each sample was variable. Most variability occurred in or around the D-gene segment, which could be due to splicing, N- and P-nucleotide additions, and deletions during somatic recombination. D-gene usage also appeared to be a factor determining CDR3 length. This was evidenced by decreasing D-gene representation across CDR3 selections of decreasing length compared to the relative conservation of the V- and J-gene segments, though J-gene conservation seemed to decrease among extremely short CDR3s. Overall, the V-gene segment appeared to remain the most uniform.

While Ig $\kappa$  contains mostly CDR3 sequences that are nine amino acids in length, many highly abundant VJ-gene segment combinations (such as V110 and J2) contained CDR3s of multiple lengths. Unlike IgH, the alignments of VJ pairings were relatively uniform among Ig $\kappa$  as

compared to the germline sequence in CDR3 sequences that were eight, nine and ten amino acids in length (Figure 3.7D).

## Discussion

To our knowledge, these data are the first unamplified sampling of the mouse antibody repertoire that has been described. Others have looked at Ig-gene segment usage with other strategies (26, 39-42) but we wanted to determine if a straight-forward RNA-Seq approach would provide us with a reasonable assessment of B-cell Ig-segment use without the limitations that amplification methods introduce.

To minimize potential single animal aberrations and repertoire skewing, we pooled splenic tissue of four unimmunized mice in three biological replicates. This approach was successful since we saw less variation with pooled samples compared to data sets that are made up of single mice (shown in chapter four). Grieff *et al.* demonstrated that CDR3 and VDJ composition in pooled mouse samples was less polarized than that of an individually sequenced mice subjected to antigen challenge (43). Therefore, our data are consistent with that study.

The most common VH-gene segment was V1-80, which was the most common in pools one and two. V6-3 was the most common in pool three. In selecting the ten most common VH-gene segments from each pool, we identified 14 different VH-gene segments, with heavy overlap among pools. All VH-gene segments isolated comprised between 1% to 8% of the repertoire. We also saw that V $\kappa$ -gene segment usage was comparable among mouse pools, with 16 gene segments comprising between 1.3% and 10.1% of the repertoire. Although we have found that gene segment use among all three pooled sample groups was relatively similar, there were some differences. V5-39 was observed at a high frequency in pool one (10.1%) as compared to pools two (1.3%) and three (3.5%). This skewing could be the result of a mouse within pool one

responding to a specific antigen that other mice did not respond to or it can represent the natural variability of mice and the randomness of antibody gene selection and rearrangement and a pool of four individuals still has relative uniqueness. Studies on humans have revealed similar overlap. The percentages of V(D)J usage is similar among individuals, in spite of them being outbred populations (27).

Some have suggested that V-gene segment usage may be skewed (44). Chi-square analyses of VH- and V $\kappa$ -gene segment use in our data set would support this contention since several VH and V $\kappa$ -gene segments were used more frequently than expected. Even though we have analyzed three independent biological samples made up of 4 mouse pools, we recognize that an even larger data set will be needed to conclusively settle this discussion. Additional studies looking at epigenetic changes or other transcriptional regulatory elements such as the characterization performed by Choi et al. (45) might also help understand mechanisms of V-gene segment selection.

For the D-gene segment, three usage levels were clearly detected. D1-1 was the most used gene segment in all three pools; comprising around 26% of the total repertoire. Over 30% of D-gene segments could not be identified, likely due to the short length of the D-gene segment. Nevertheless, we do see different populations of antibodies even when they do share similar VDJ-genes segments. Some have large D-gene segments where others have little recognizable sequence.

JH-gene segment usage was again, relatively uniform, J2 was the most common among all three pools, with over 32% use in the repertoire. J1, J3, and J4 were evenly represented among all three polls totaling between 18-27% of the repertoire. Gene segments with less than six nucleotides were unable to be identified and occur at less than 0.1% of the heavy-chain repertoire. J $\kappa$ -gene segment usage is somewhat evenly distributed among J1, J2, and J5 comprising between 25-30%

of the repertoire, in agreement with the findings of Aoki-ota *et al.* (44). Lu *et al.* found that a slightly different J $\kappa$  expression profile, possibly reflecting strain specific usage of J $\kappa$  (26). As paralleled in the heavy-chain data, gene segments with less than six nucleotides were rare; occurring in less than three percent of the total repertoire.

Constant region usage in the heavy-chain was heavily dominated by IgM, which reflects the “naïve” status of our mice. Although IgM comprised over 78% of the total identified constant regions we did see the expression of IgG, IgA and IgD. IgE was rare, being detected in only pool 3. However, when compared to serum data, even in naive mice, there is a high level of circulating IgG (46). This could be due to a large B-cell population in the spleen that is not secreting antibody into the bloodstream.

We looked at the common H-CD3 sequences among the three mouse pools. There was little overlap; with only 75 H-CDR3s detected in all three pools and between 92 and 163 common when we just looked at two pools. We detected between 4.6k to 6.2k unique sequences found only in each respective pool. While we sampled a small fraction of CDR3s present in the total antibody repertoire Lu *et al.* used primer amplification to enrich for IgH transcripts and still found high CDR3 variability among individuals (26). Similarly, in a study comparing monozygotic twins, Glanville *et al.*, also demonstrated that CDR3 profiles between the individuals were quite diverse despite similar gene family usage between the twins (47).

When we examined the  $\kappa$ -CDR3 usage among the three biological samples, there was a higher proportion of common  $\kappa$ -CDR3s [436]. Unique  $\kappa$ -CDR3s within each pool ranged from 688 to 709 CDR3 identified within all three pools and between 100 and 143 CDR3 identified in only two pools. One explanation for light-chain CDR3 length homogeneity may be selection due to light-chain editing that occurs during B-cell maturation.

The small numbers of overlapping CDR3 sequences among our three pooled samples suggests that significant variation in the idiotypes could develop, even within an inbred population of mice. We were curious if the size of the total pool of B cells could be estimated from our observational data. Using a model of capture-recapture methodology (48), using the Chapman estimator (49) and the number of common heavy-chain CDR3 sequences seen in each of our samplings, we estimated our B-cell pool to range from 1.5-14 x 10<sup>6</sup> cells. If we assume that there are 3.5 x 10<sup>6</sup> B cells in a nine to eleven week-old female C57BL/6J mouse spleen, and our mouse pools were made up of 4 spleens, this estimate of the possible B-cell pool is reasonably accurate, especially if we take into account some of the CDR3 sequences were detected multiple times (multiple B cells with the same IgH).

While CDR3 is commonly used to describe the antibody repertoire, many studies have reported the combinations of the V(D)J (20, 43, 50-52). Compiling CDR3 nucleotide alignments allowed us to visualize the significance of individual gene segment involvement with the CDR3 in the context of specific V(D)J combinations. Sequencing outside of CDR3 also reveals biologically relevant information about antigen binding and allows for further characterization and potential lineage determination. Our sequencing technique enabled the identification of V(D)J-gene segments in addition to constant region, providing insight into pairing of V-gene families with (D)J-gene segments.

The current data reinforces the unique generation of B-cell diversity even in populations that one would expect to be similar. The diversity in H-CDR3 sequences among the three biological samples is paralleled in humans (27, 47). The information about the unchallenged Ig-gene repertoire also has other uses. It provides a comparative foundation when looking at host response to antigen (53). Ig-gene sequences have been used to isolate therapeutic antibodies for

influenza in a mouse model and were a valuable tool in the detection of antigen specific responses (16, 34).

Amplification with multiplex V-gene segment primers can result in bias within the HTS dataset as some primers may hybridize more efficiently with certain V-gene segments (54). While a lack of amplification may extricate primer bias, we knew that it might come at a cost of potentially excluding rare B-cell clones. In humans, for example, a single clone may only comprise 0.1% to 0.3% of the repertoire (55). We have explored the differences between samples that have and have not been amplified and found a moderate correlation ( $R^2 = 0.5815, 0.5855, p < 0.0001$ , Rettig et al., manuscript in preparation). While some of the differences arose from expected depth-of-sequencing issues, we unexpectedly found that discrepancies also resulted from gene segments being detected in the unamplified data set not detected in the amplified data sets (Rettig et al. manuscript in preparation).

Another issue which may affect the data stems from the use of whole spleen tissue rather than isolated B-cell populations (34). Although our approach was necessary to accommodate requirements of a separate investigation (35), we are aware that the inclusion of extraneous cells as result of using whole tissue could reduce the recovery of rare B-cell clones. In addition, some bias might be introduced because of cell subpopulation stability and frequency in whole spleen tissue (56, 57). In spite of the limitations of our methodology, it appears that the repertoire we detected correlated with mouse studies that have used selection and amplification methods of various kinds. For example, Collins *et al.* detected five of the same VH genes that we detected among our highest 10 used VH-gene segments. JH2 was also the most frequently detected in both of our studies (33). Yang and Kaplinski detected V-gene segment use that paralleled our findings with V1-26 identified by them as the most frequently used (11, 40).

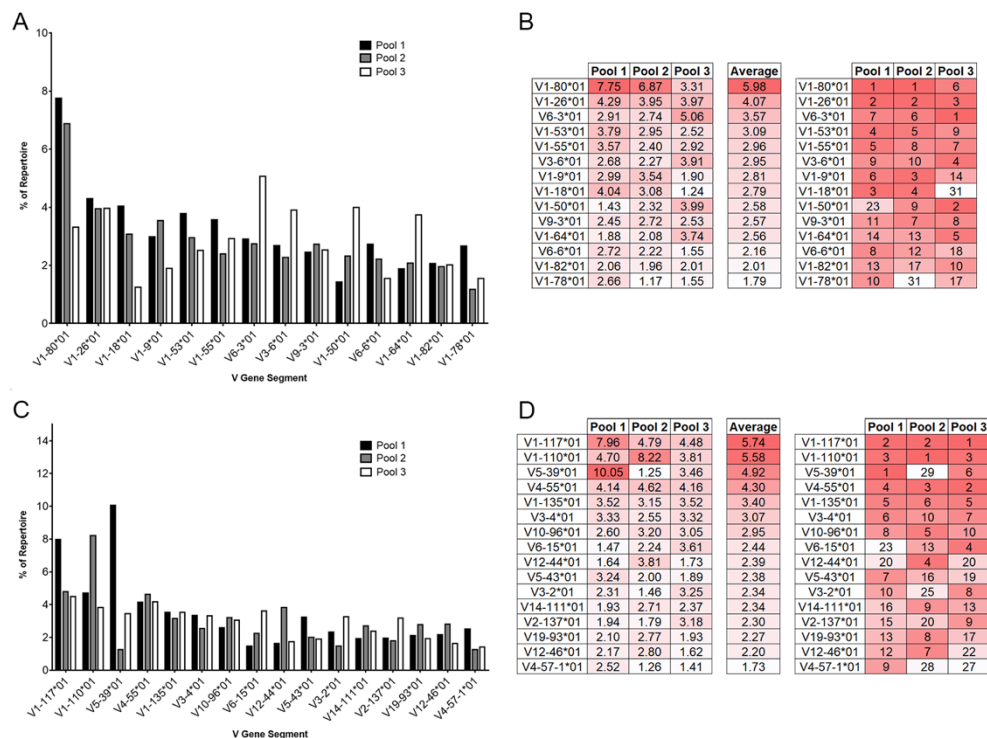


Few studies have explored the light chain repertoire, however, more characterization will be possible with increasing use of single cell amplification (58-60). While strain specificity has been reported (33, 41), many V $\kappa$ -gene segments that were represented over one-percent of the time in unimmunized BALB/c mice were also identified in our study (26). Aoki-Ota *et al.* also noted V $\kappa$ -gene segment skewing in their assessment of unimmunized C57BL/6 mice (44). These similarities also suggest that the lack of amplification did not dramatically affect our assessment of the B-cell repertoire, and the differences seen are likely due to mouse-to-mouse variation that still manifests in our pooled samples.

In conclusion, we have presented an unamplified view of the conventionally housed, unimmunized, antibody repertoire. We lay the foundation for future work in our lab to characterize the unamplified whole tissue repertoire of the C57BL/6 mouse.

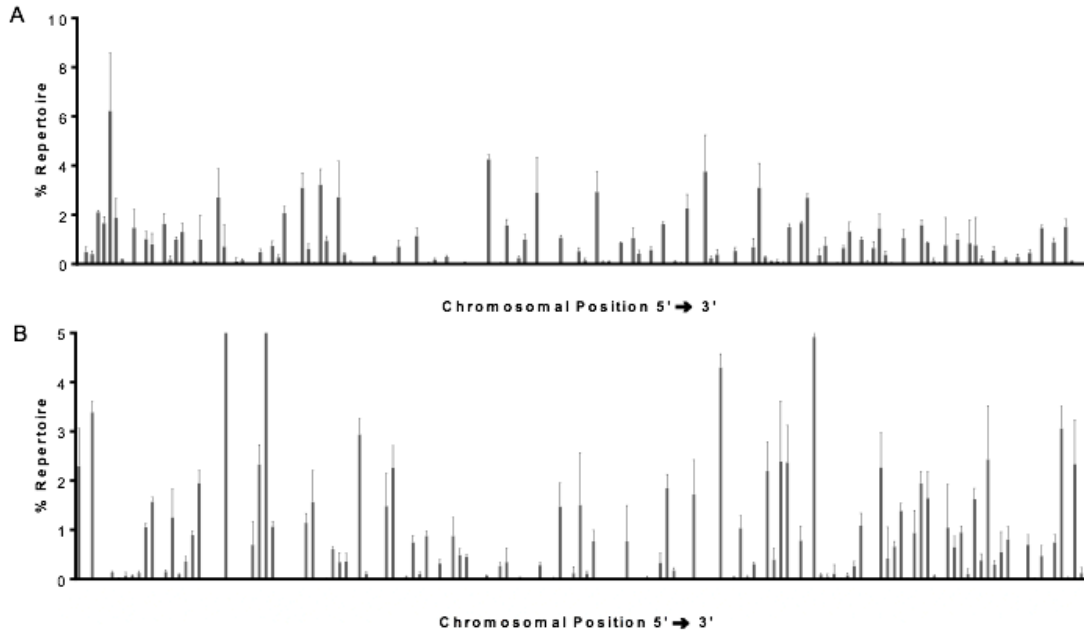
## Figures and Tables

Figure 3.1 V-gene segment usage among unimmunized mouse pool



Sequencing reads mapped to each individual gene segment were divided by the total sequencing reads of all identified gene segments from each mouse pool for a normalized comparison between pools. (A) The VH representing the ten most abundant gene segments from each mouse pool are displayed. (B) The rankings of each gene segment contained within the top 10 most abundant VH from at least one of the mouse pools are compared. The most abundant gene segment is ranked as 1. Dark red indicates higher rank moving to white, of lower rank. Similarly, the top 10 abundant V $\kappa$  are displayed (C-D).

**Figure 3.2 V-gene segment usage by chromosomal location**



V-gene segment usage among unimmunized mouse pools for IgH (A) and Igκ (B) by chromosomal location. Gene segments are shown in order of chromosomal position (5' to 3'). The average value from three mouse pools for each CDR3 length is shown. Distribution was assessed via Chi-square analysis in R (version 3.3.1) ( $p < 0.0001$ ).

**Figure 3.3 D-, J-gene segment and constant region usage**

**A**

	Pool 1	Pool 2	Pool 3	Average		Pool 1	Pool 2	Pool 3
Undeter	34.75	31.64	31.93	32.77	Undeter	1	1	1
D1-1*01	27.57	25.89	26.01	26.49	D1-1*01	2	2	2
D2-3*01	6.47	15.47	9.62	10.52	D2-3*01	6	3	4
D2-4*01	8.42	7.66	11.57	9.22	D2-4*01	4	5	3
D4-1*01	8.43	7.79	8.28	8.17	D4-1*01	3	4	5
D2-5*01	7.99	5.51	6.77	6.75	D2-5*01	5	6	6
D3-2*02	2.77	2.23	2.69	2.57	D3-2*02	7	8	7
D3-1*01	2.19	2.80	1.92	2.30	D3-1*01	8	7	8
D6-2*02	0.41	0.31	0.55	0.42	D6-2*02	10	10	9
D5-1*01	0.37	0.51	0.37	0.42	D5-1*01	11	9	10
D5-5*01	0.62	0.18	0.28	0.36	D5-5*01	9	11	11

**B**

	Pool 1	Pool 2	Pool 3	Average		Pool 1	Pool 2	Pool 3
J2*01	36.47	31.95	37.97	35.46	J2*01	1	1	1
J3*01	19.47	27.27	19.82	22.19	J3*01	4	2	3
J4*01	19.77	22.13	22.70	21.54	J4*01	3	3	2
J1*03	24.25	18.62	19.44	20.77	J1*03	2	4	4
Undeter	0.03	0.04	0.07	0.05	Undeter	5	5	5

**C**

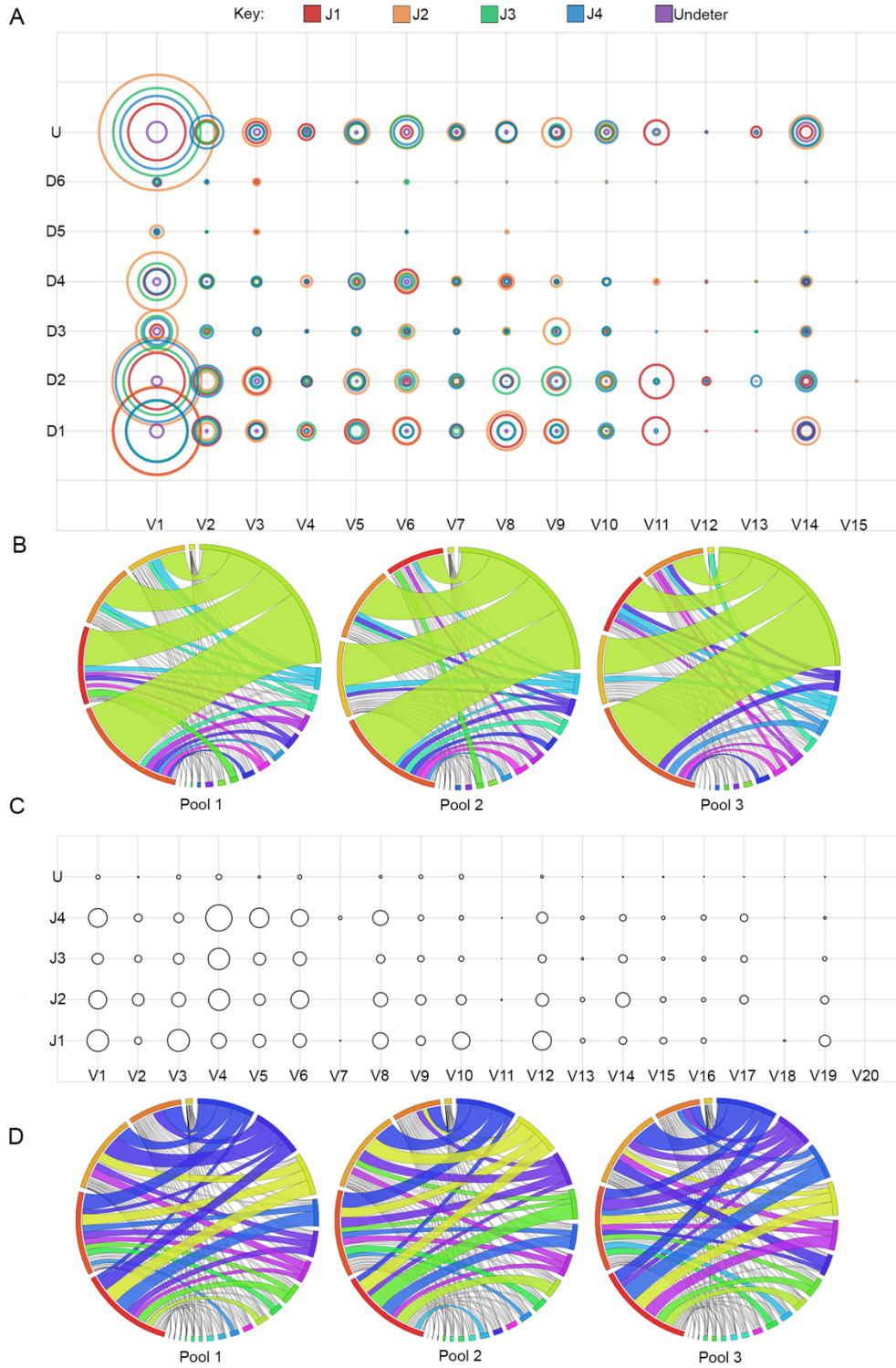
	Pool 1	Pool 2	Pool 3	Average		Pool 1	Pool 2	Pool 3
IgM	83.63	82.23	78.12	81.33	IgM	1	1	1
IgG	7.38	10.94	10.62	9.64	IgG	2	2	2
IgD	3.81	5.02	4.98	4.60	IgD	4	3	4
IgA	5.17	1.82	6.27	4.42	IgA	3	4	3
IgE	0	0	0.02	0.01	IgE	5	5	5

**D**

	Pool 1	Pool 2	Pool 3	Average		Pool 1	Pool 2	Pool 3
J1*01	29.57	29.12	30.81	29.83	J1*01	1	1	1
J2*01	26.27	27.29	26.31	26.62	J2*01	2	2	3
J5*01	24.92	26.18	26.33	25.81	J5*01	3	3	2
J4*01	17.09	15.29	14.12	15.50	J4*01	4	4	4
Undeter	2.15	1.66	2.43	2.08	Undeter	5	5	5

Percent abundance of IgH D- (A) and J- (B) gene segments, IgH constant regions (C) and Igk J-gene segments (D). Sequencing reads corresponding to each gene segment or constant region were divided by the total number of gene segments or constant regions identified in each mouse pool for normalized comparison between pools. The most abundant gene segment is ranked as 1. Dark red indicates higher rank moving to white, of lower rank. Sequencing reads designated undetermined (undeter) where portions of a D- or J-gene segment were identified but unable to be assigned to a specific C57BL/6J D- or J-gene segment.

**Figure 3.4 V(D)J-gene segment combinations**



Combinations of V-gene families with DJ-gene segments for IgH (A) and J-gene segments for Igκ (C). Increasing pairing frequency of V(D)J is represented by larger circles. Sequencing reads in which more than one C57BL/6 J-gene segment was attributed or too few nucleotides were present in the J-gene segment for designation by IMGT have been classified as undetermined (U). Pairing frequency is also represented by Circos graphs for IgH (B) and Igκ (D). Circos Plot Labels (starting at 12:00 position and the largest arc and continuing clockwise)

B – Pool 1: V1, V2, V14, V8, V6, V3, V9, V5, V11, V10, V7, V4, V13, V12, V15, J2, J1, J3, J4, Undetermined

B – Pool 2: V1, V2, V6, V14, V5, V8, V9, V3, V10, V11, V7, V4, V12, V13, V15, J2, J4, J3, J1, Undetermined

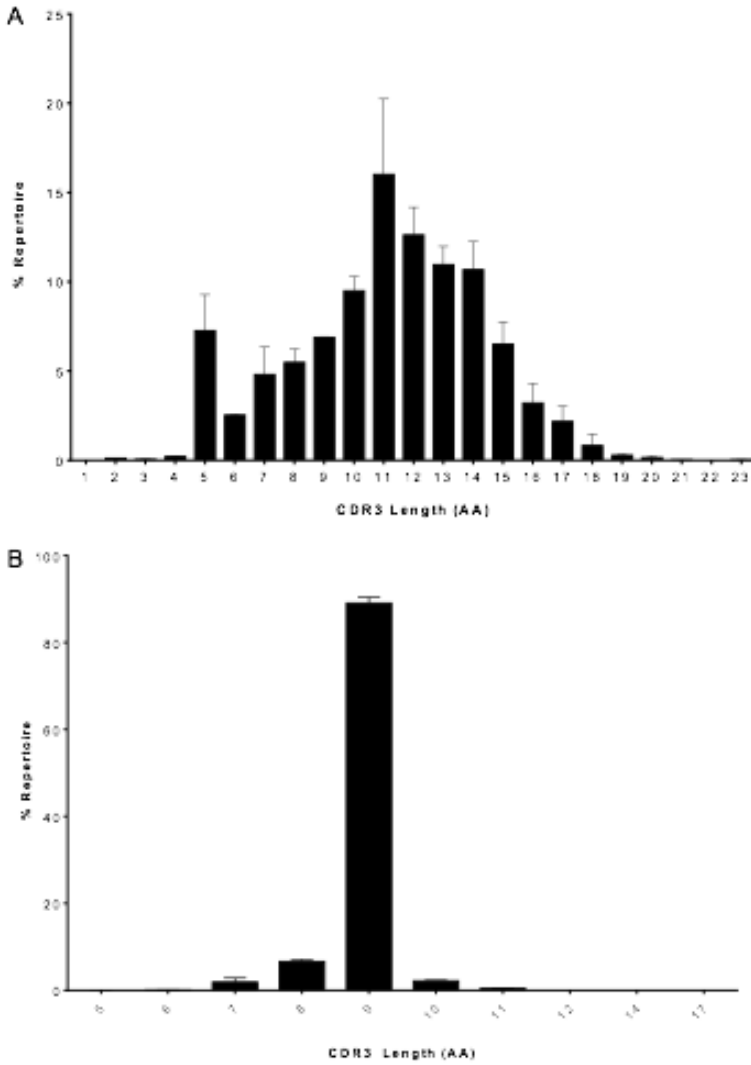
B – Pool 3: V1, V6, V2, V3, V14, V8, V9, V5, V10, V7, V11, V4, V13, V15, V12, J2, J4, J1, J3, Undetermined

D – Pool 1: V4, V5, V1, V3, V6, V8, V12, V14, V10, V9, V2, V19, V17, V16, V15, V13, V18, V7, V11, J1, J2, J5, J4, Undetermined

D – Pool 2: V4, V1, V6, V12, V3, V8, V10, V14, V2, V9, V5, V19, V17, V15, V16, V13, V7, V11, V18, J1, J2, J5, J4, Undetermined

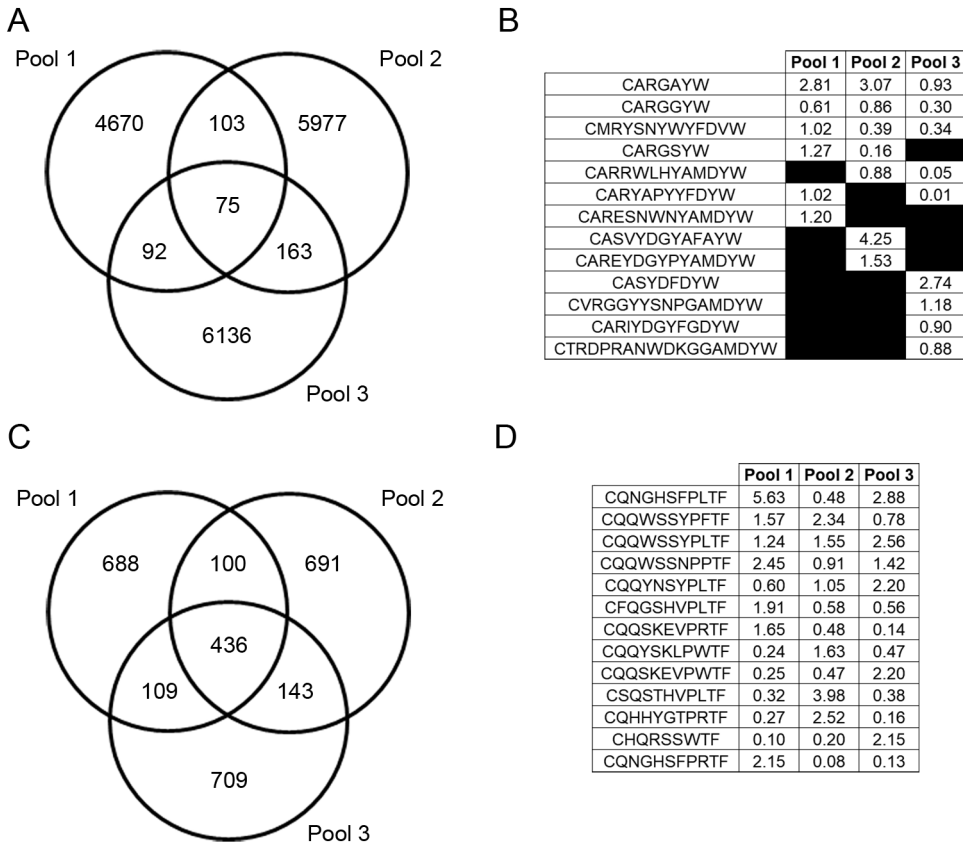
D – Pool 3: V4, V6, V3, V1, V8, V5, V12, V10, V2, V14, V9, V17, V19, V16, V13, V15, V7, V11, V20, J1, J2, J5, J4, Undetermined

**Figure 3.5 CDR3 length**



CDR3 length for IgH (A) and Igk (B). The average percent of repertoire of each CDR3 amino acid length from three mouse pools is displayed.

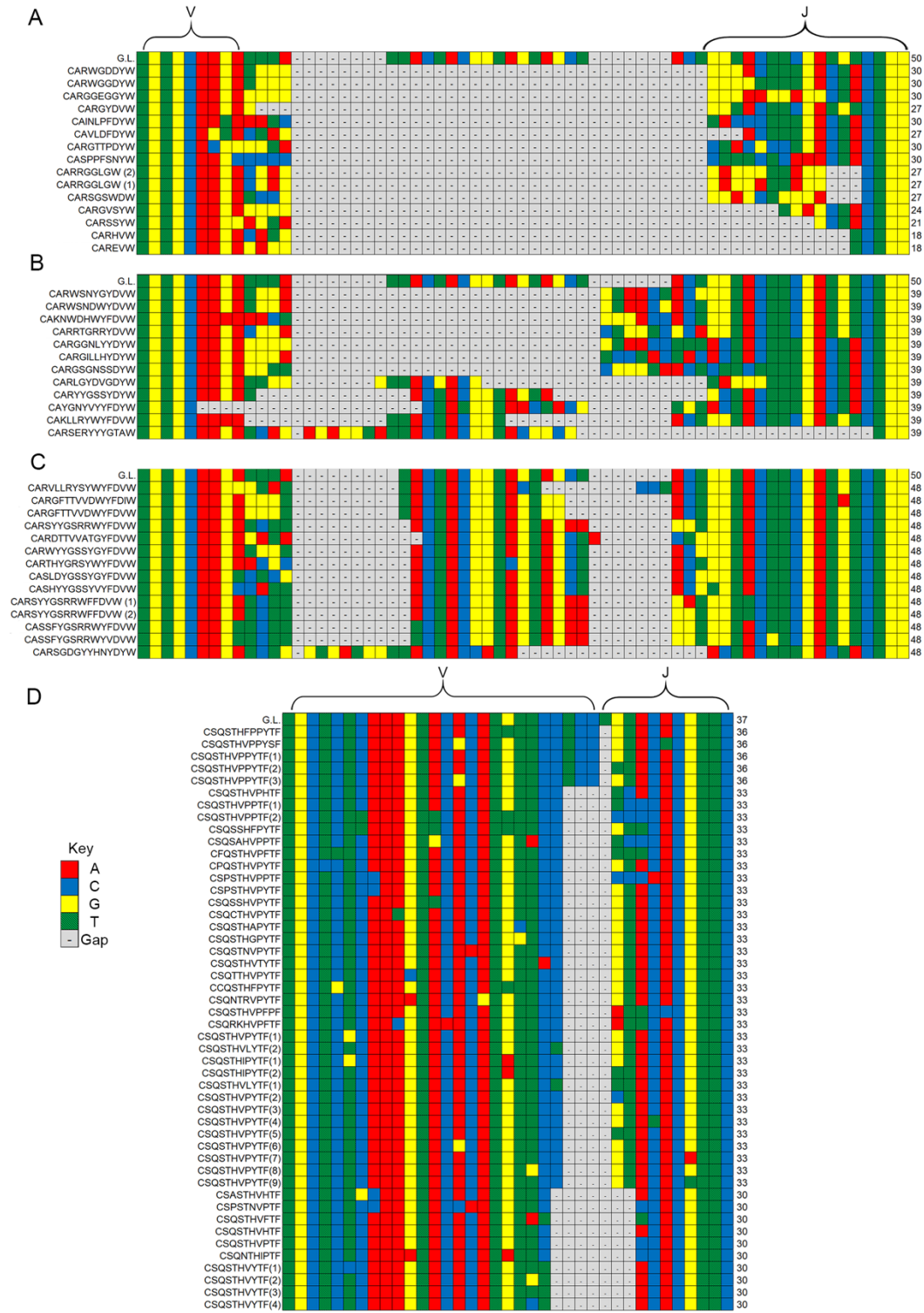
**Figure 3.6 Overlap of unique CDR3 sequences within pools and top CDR3 sequences**



A Venn diagram displays the overlap of the number of unique CDR3 amino acid sequences among mouse pools for IgH (A) and Igκ (C). The percent of repertoire for the top 5 CDR3 amino acid from each mouse pool are shown for IgH (B) and Igκ (D).



**Figure 3.7 Alignments of top V(D)J-gene segment combinations**



Comparison of CDR3 alignments in gene segment combinations (IGHV1-26, IGHD1-1, IGHJ1) coding for a predominantly short (A), median length (B), and long (C) H-CDR3 region and  $\kappa$ -CDR3 (IG $\kappa$ V1-110, IG $\kappa$ J-2) (D). The germline nucleotide (G.L.) sequence is identified at the top of each alignment. Each nucleotide sequence is labeled with its corresponding amino acid sequence. Nucleotide sequences coding for identical amino acid sequences are labeled with numbers (1, 2, 3, etc.) corresponding with the alignment order. The V and J-gene segments for each alignment are labeled, however due to the variability in the D-gene segment it is not bracketed, but is identifiable by the germline sequence provided.

**Table 3.1 Sequencing and mapping statistics from mouse pools 1, 2, and 3.**

	Pool 1	Pool 2	Pool 3
Total Reads	25.1 M <sup>a</sup>	31.4 M	32.7 M
Post Cleaning	12.0 M	30.9 M	32.0 M
Productive IgH	8714	11200	10224
Unknown IgH	14271	27896	18756
Productive Igκ	12199	15111	13433
Unknown Igκ	13264	36741	34559

<sup>a</sup>M= Million sequencing reads

## References

1. Shahaf, G., M. Barak, N. S. Zuckerman, N. Swerdlin, M. Gorfine, and R. Mehr. 2008. Antigen-driven selection in germinal centers as reflected by the shape characteristics of immunoglobulin gene lineage trees: a large-scale simulation study. *J. Theor. Biol.* 255: 210-222.
2. Cory, S. 2015. Masterminding B Cells. *J. Immunol.* 195: 763-765.
3. Hozumi, N., and S. Tonegawa. 1976. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl. Acad. Sci.* 73: 3628-3632.
4. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575-581.
5. Early, P., H. Huang, M. Davis, K. Calame, and L. Hood. 1980. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. *Cell* 19: 981-992.
6. Tonegawa, S. 1976. Reiteration frequency of immunoglobulin light chain genes: further evidence for somatic generation of antibody diversity. *Proc. Natl. Acad. Sci.* 73: 203-207.
7. Kabat, E. A., T. T. Wu, and H. Bilofsky. 1979. Evidence supporting somatic assembly of the DNA segments (minigenes), coding for the framework, and complementarity-determining segments of immunoglobulin variable regions. *J. Exp. Med.* 149: 1299-1313.
8. Xu, Z., H. Zan, E. J. Pone, T. Mai, and P. Casali. 2012. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. *Nat. Rev. Immunol.* 12: 517-531.
9. Ippolito, G. C., R. L. Schelonka, M. Zemlin, Ivanov, II, R. Kobayashi, C. Zemlin, G. L. Gartland, L. Nitschke, J. Pelkonen, K. Fujihashi, K. Rajewsky, and H. W. Schroeder, Jr.

2006. Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. *J. Exp. Med.* 203: 1567-1578.
10. Greiff, V., E. Miho, U. Menzel, and S. T. Reddy. 2015. Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. *Trends Immunol.* 36: 738-749.
  11. Yang, Y., C. Wang, Q. Yang, A. B. Kantor, H. Chu, E. E. Ghosn, G. Qin, S. K. Mazmanian, J. Han, and L. A. Herzenberg. 2015. Distinct mechanisms define murine B cell lineage immunoglobulin heavy chain (IgH) repertoires. *eLife* 4: e09083.
  12. Jiang, N., J. A. Weinstein, L. Penland, R. A. White, 3rd, D. S. Fisher, and S. R. Quake. 2011. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci.* 108: 5348-5353.
  13. Racanelli, V., D. Sansonno, C. Piccoli, F. P. D'Amore, F. A. Tucci, and F. Dammacco. 2001. Molecular characterization of B cell clonal expansions in the liver of chronically hepatitis C virus-infected patients. *J. Immunol.* 167: 21-29.
  14. Parameswaran, P., Y. Liu, K. M. Roskin, K. K. Jackson, V. P. Dixit, J. Y. Lee, K. L. Artiles, S. Zompi, M. J. Vargas, B. B. Simen, B. Hanczaruk, K. R. McGowan, M. A. Tariq, N. Pourmand, D. Koller, A. Balmaseda, S. D. Boyd, E. Harris, and A. Z. Fire. 2013. Convergent antibody signatures in human dengue. *Cell Host Microbe* 13: 691-700.
  15. Reddy, S. T., X. Ge, A. E. Miklos, R. A. Hughes, S. H. Kang, K. H. Hoi, C. Chrysostomou, S. P. Hunicke-Smith, B. L. Iverson, P. W. Tucker, A. D. Ellington, and G. Georgiou. 2010. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature Biotechnol.* 28: 965-969.

16. Gray, S. A., M. Moore, E. J. Vandenberg, R. P. Roque, R. A. Bowen, N. Van Hoven, S. R. Wiley, and C. H. Clegg. 2016. Selection of therapeutic H5N1 monoclonal antibodies following IgVH repertoire analysis in mice. *Antivir. Res.* 131: 100-108.
17. Galson, J. D., A. J. Pollard, J. Truck, and D. F. Kelly. 2014. Studying the antibody repertoire after vaccination: practical applications. *Trends Immunol.* 35: 319-331.
18. Ademokun, A., Y. C. Wu, V. Martin, R. Mitra, U. Sack, H. Baxendale, D. Kipling, and D. K. Dunn-Walters. 2011. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* 10: 922-930.
19. Rosenquist, R., U. Thunberg, A. H. Li, E. Forestier, G. Lonnerholm, J. Lindh, C. Sundstrom, J. Sallstrom, D. Holmberg, and G. Roos. 1999. Clonal evolution as judged by immunoglobulin heavy chain gene rearrangements in relapsing precursor-B acute lymphoblastic leukemia. *Eur. J. Haematol.* 63: 171-179.
20. Bashford-Rogers, R. J., A. L. Palser, B. J. Huntly, R. Rance, G. S. Vassiliou, G. A. Follows, and P. Kellam. 2013. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res.* 23: 1874-1884.
21. Bashford-Rogers, R. J., K. A. Nicolaou, J. Bartram, N. J. Goulden, L. Loizou, L. Koumas, J. Chi, M. Hubank, P. Kellam, P. A. Costeas, and G. S. Vassiliou. 2016. Eye on the B-ALL: B-cell receptor repertoires reveal persistence of numerous B-lymphoblastic leukemia subclones from diagnosis to relapse. *Leukemia* 30: 2312-2321.
22. van Belzen, N., P. E. Hupkes, D. Doekharan, M. Hoogeveen-Westerveld, L. C. Dorsers, and M. B. van't Veer. 1997. Detection of minimal disease using rearranged immunoglobulin heavy chain genes from intermediate- and high-grade malignant B cell non-Hodgkins lymphoma. *Leukemia* 11: 1742-1752.

23. Zuckerman, N. S., W. A. Howard, J. Bismuth, K. Gibson, H. Edelman, S. Berrih-Aknin, D. Dunn-Walters, and R. Mehr. 2010. Ectopic GC in the thymus of myasthenia gravis patients show characteristics of normal GC. *Eur. J. Immunol.* 40: 1150-1161.
24. Tan, Y. C., S. Kongpachith, L. K. Blum, C. H. Ju, L. J. Lahey, D. R. Lu, X. Cai, C. A. Wagner, T. M. Lindstrom, J. Sokolove, and W. H. Robinson. 2014. Barcode-enabled sequencing of plasmablast antibody repertoires in rheumatoid arthritis. *Arthritis Rheumatol. (Hoboken, N.J.)* 66: 2706-2715.
25. Greiff, V., P. Bhat, S. C. Cook, U. Menzel, W. Kang, and S. T. Reddy. 2015. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* 7: 49.
26. Lu, J., T. Panavas, K. Thys, J. Aerssens, M. Naso, J. Fisher, M. Ryczyn, and R. W. Sweet. 2014. IgG variable region and VH CDR3 diversity in unimmunized mice analyzed by massively parallel sequencing. *Mol. Immunol.* 57: 274-283.
27. Briney, B. S., J. R. Willis, B. A. McKinney, and J. E. Crowe, Jr. 2012. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. *Genes Immun.* 13: 469-473.
28. Boyd, S. D., B. A. Gaeta, K. J. Jackson, A. Z. Fire, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, and A. M. Collins. 2010. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* 184: 6986-6992.
29. Watson, C. T., K. M. Steinberg, J. Huddleston, R. L. Warren, M. Malig, J. Schein, A. J. Willsey, J. B. Joy, J. K. Scott, T. A. Graves, R. K. Wilson, R. A. Holt, E. E. Eichler, and

- F. Breden. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.* 92: 530-546.
30. Sasso, E. H., K. W. Van Dijk, and E. C. Milner. 1990. Prevalence and polymorphism of human VH3 genes. *J. Immunol.* 145: 2751-2757.
31. Milner, E. C., W. O. Hufnagle, A. M. Glas, I. Suzuki, and C. Alexander. 1995. Polymorphism and utilization of human VH Genes. *Ann. N. Y. Acad. Sci.* 764: 50-61.
32. Wang, Y., K. J. Jackson, B. Gaeta, W. Pomat, P. Siba, W. A. Sewell, and A. M. Collins. 2011. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* 63: 259-265.
33. Collins, A. M., Y. Wang, K. M. Roskin, C. P. Marquis, and K. J. Jackson. 2015. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Phil. Trans. R. Soc. Lon. B* 370.
34. Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnol.* 32: 158-168.
35. Rettig, T. A., C. Ward, M. J. Pecaut, and S. K. Chapes. 2017. Validation of Methods to Assess the Immunoglobulin Gene Repertoire in Tissues Obtained from Mice on the International Space Station. *Gravit. Space Res.* 5:2-23.
36. Alamyar, E., P. Duroux, M. P. Lefranc, and V. Giudicelli. 2012. IMGT((R)) 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.* 882: 569-604.



37. Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30: 3059-3066.
38. Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res.* 19: 1639-1645.
39. de Bono, B., M. Madera, and C. Chothia. 2004. VH gene segments in the mouse and human genomes. *J. Mol. Biol.* 342: 131-143.
40. Kaplinsky, J., A. Li, A. Sun, M. Coffre, S. B. Koralov, and R. Arnaout. 2014. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc. Natl. Acad. Sci.* 111: E2622-2629.
41. Greiff, V., U. Menzel, E. Miho, C. Weber, R. Riedel, S. Cook, A. Valai, T. Lopes, A. Radbruch, T. H. Winkler, and S. T. Reddy. 2017. Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep.* 19: 1467-1478.
42. Kono, N., L. Sun, H. Toh, T. Shimizu, H. Xue, O. Numata, M. Ato, K. Ohnishi, and S. Itamura. 2017. Deciphering antigen-responding antibody repertoires by using next-generation sequencing and confirming them through antibody-gene synthesis. *Biochem. Biophys. Res. Commun.* 487: 300-306.
43. Greiff, V., U. Menzel, U. Haessler, S. C. Cook, S. Friedensohn, T. A. Khan, M. Pogson, I. Hellmann, and S. T. Reddy. 2014. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.* 15: 40.

44. Aoki-Ota, M., A. Torkamani, T. Ota, N. Schork, and D. Nemazee. 2012. Skewed primary Igkappa repertoire and V-J joining in C57BL/6 mice: implications for recombination accessibility and receptor editing. *J. Immunol.* 188: 2305-2315.
45. Choi, N. M., S. Loguercio, J. Verma-Gaur, S. C. Degner, A. Torkamani, A. I. Su, E. M. Oltz, M. Artyomov, and A. J. Feeney. 2013. Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom V gene rearrangement frequencies. *J. Immunol.* 191: 2393-2402.
46. Klein-Schneegans, A. S., L. Kuntz, P. Fonteneau, and F. Loor. 1989. Serum concentrations of IgM, IgG1, IgG2b, IgG3 and IgA in C57BL/6 mice and their congenics at the *lpr* (lymphoproliferation) locus. *J. Autoimmun.* 2: 869-875.
47. Glanville, J., T. C. Kuo, H. C. von Budingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G. R. Mehta, J. R. Oksenberg, S. L. Hauser, D. R. Cox, A. Rajpal, and J. Pons. 2011. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci.* 108: 20066-20071.
48. Chao, A., P. K. Tsay, S. H. Lin, W. Y. Shau, and D. Y. Chao. 2001. The applications of capture-recapture models to epidemiological data. *Statist. Med.* 20: 3123-3157.
49. Chapman, D. G., and B. University of California. 1951. *Some properties of the hypergeometric distribution with applications to zoological sample censuses*. University of California Press, Berkeley.
50. Calis, J. J., and B. R. Rosenberg. 2014. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol.* 35: 581-590.
51. Kunik, V., B. Peters, and Y. Ofran. 2012. Structural consensus among antibodies defines the antigen binding site. *PLoS Comput. Biol.* 8: e1002388.

52. Sela-Culang, I., V. Kunik, and Y. Ofran. 2013. The structural basis of antibody-antigen recognition. *Front. Immunol.* 4: 302.
53. Banga, S., J. D. Coursen, S. Portugal, T. M. Tran, L. Hancox, A. Ongoiba, B. Traore, O. K. Doumbo, C. Y. Huang, J. T. Harty, and P. D. Crompton. 2015. Impact of acute malaria on pre-existing antibodies to viral and vaccine antigens in mice and humans. *PloS one* 10: e0125090.
54. Wardemann, H. B., C.E. 2017. Novel Approaches to Analyze Immunoglobulin Repertoires. *Trends Immunol.* 38: 471-482.
55. Boyd, S. D., E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, and A. Z. Fire. 2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* 1: 12ra23.
56. Allman, D. M., S. E. Ferguson, V. M. Lentz, and M. P. Cancro. 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.
57. Allman, D., R. C. Lindsley, W. DeMuth, K. Rudd, S. A. Shinton, and R. R. Hardy. 2001. Resolution of three nonproliferative immature splenic B cell subsets reveals multiple selection points during peripheral B cell maturation. *J. Immunol.* 167: 6834-6840.
58. DeKosky, B. J., G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke, T. Dorner, S. F. Andrews, P. C. Wilson, S. P. Hunicke-Smith, C. G. Willson, A. D. Ellington, and G. Georgiou. 2013. High-throughput sequencing of the

- paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnol.* 31: 166-169.
59. DeKosky, B. J., T. Kojima, A. Rodin, W. Charab, G. C. Ippolito, A. D. Ellington, and G. Georgiou. 2015. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nature Med.* 21: 86-91.
60. Busse, C. E., I. Czogiel, P. Braun, P. F. Arndt, and H. Wardemann. 2014. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J. Immunol.* 44: 597-603.

# **Chapter 4 - Effects of Spaceflight on the Antibody Repertoire of Unimmunized C57BL/6 Mice**

## **Abstract**

### **Abstract**

Spaceflight has been shown to suppress the adaptive immune response, altering the distribution and function of lymphocyte populations. B lymphocytes express highly specific and highly diversified receptors, known as immunoglobulins (Ig), that directly bind and neutralize pathogens. Ig diversity is achieved through the enzymatic splicing of gene segments within the genomic DNA of each B cell in a host. The collection of Ig specificities within a host, or Ig repertoire, has been increasingly characterized in both basic research and clinical settings using high-throughput sequencing technology (HTS). We utilized HTS to test the hypothesis that spaceflight affects the B-cell repertoire. To test this hypothesis, we characterized the impact of spaceflight on the unimmunized Ig repertoire of C57BL/6 mice that were flown aboard the International Space Station (ISS) during the Rodent Research One validation flight in comparison to ground controls. Individual gene segment usage was similar between ground control and flight animals, however, gene segment combinations and the junctions in which gene segments combine was varied among animals within and between treatment groups. We also found that spontaneous somatic mutations in the IgH and Igk gene loci were not increased. These data suggest that space flight did not affect the B cell repertoire of mice flown and housed on the ISS over a short period of time.

## Introduction

Spaceflight presents a unique set of challenges to the immune system. For example, spaceflight alters T- and B-lymphocyte functions, including recall responses in astronauts aboard the space shuttle and cytokine responses after missions to the international space station (ISS) (1-8). In addition to functional changes, lymphocyte subpopulations are altered. CD8<sup>+</sup> T-cell numbers were increased during flight while other T-cell subsets were decreased (2). Similar changes in phenotype also occur in animal and tissue culture systems during spaceflight or ground-based spaceflight analogs such as anti-orthostatic suspension (AOS) (9-13).

Lymphocyte subpopulations change in response to spaceflight (14-21) and AOS (22, 23). Splenic T- and B-lymphocyte counts were decreased in mice flown on the 13-day mission of the Space Shuttle Endeavor (STS-118) compared to ground controls (16). In the AOS model, Wei et al. found a reduced number of both T and B lymphocytes in the thymus and spleen of hindlimb unloaded Balb/c mice compared to normal controls (23). Reductions in the mass of lymphoid organs has also been observed (15, 22, 24-30). Spaceflight altered the phenotype of immune cells in the bone marrow, the lymphoid organ in which hematopoiesis occurs (31), and AOS reduced the number of bone marrow B-cell progenitors (32).

While many studies have characterized T-cell response to spaceflight (15, 16, 29, 33-45), fewer studies have characterized the impact of spaceflight on B-cell populations. The characterization of B-cell receptors, known as immunoglobulins (Igs), is of particular interest due to the (IgH) and light chains, which are encoded on separate loci (46). The heavy chain locus encodes multiple Variable- (V), Diversity- (D) and Joining- (J) gene segments, while the functionally equivalent  $\kappa$  (Ig $\kappa$ ) and  $\lambda$  (Ig $\lambda$ ) light chain loci contain only V- and J-gene segments (47, 48). During early B-cell development in the bone marrow, B cells undergo recombination of

heavy and light chain Ig loci, in which only one of each V(D)J-gene segment is selected for Ig use (46, 49). Random and palindromic nucleotide insertion at splice sites adds to Ig diversity (50-52).

In the Ig molecule, complementarity determining regions (CDR) confer binding specificity. CDR1 and CDR2 are encoded entirely within the V-gene segment, while CDR3 contains a portion of the 3' end of the V-gene segment, the entire D-gene segment, and a portion of the 5' end of the J gene segment (46, 53, 54). As a result of somatic recombination, B cells collectively express individual Igs that theoretically can bind virtually any pathogen.

An individual's Ig repertoire can be characterized using high-throughput sequencing (HTS) using either genomic DNA or messenger RNA sequences isolated from B-cell populations (55-57). B cells will clonally expand after antigen-Ig receptor engagement, resulting in a higher portion of target-specific Ig receptors within the B-cell population. There have been a number of HTS-based Ig repertoire studies in human disease, ranging from infectious disease (58-61), autoimmunity (62-64), and cancer (65-69). Greiff, et al. developed a profiling framework using the Ig repertoire as an indicator of an individual's immunological status (70).

Some have explored the impact of spaceflight on Ig repertoires. *In vitro* challenge of human B cells during spaceflight resulted in lower concentrations of secreted Ig (71). There was no significant difference in pre- and post-flight Ig levels in peripheral blood of astronauts who flew aboard the ISS (8, 72, 73). These samples, however, were not taken after challenge with a specific antigen. Rats immunized intraperitoneally with sheep red blood cells prior to spaceflight produced significantly less serum IgG compared to immunized ground control animals (34).

Although some have explored Ig gene segment changes in the context of spaceflight or model analogs (74-77), little has been done to characterize the impact of spaceflight on the Ig repertoire in mice. Given that changes in B cells and Ig concentrations occur during spaceflight

conditions, we tested the hypothesis that spaceflight alters the Ig repertoire of mice flown on the ISS. We examined individual Ig gene segment usage, gene segment combinations, CDR3 composition, and frame work and CDR mutations in 35-week-old, unimmunized, female C57BL/6Tac mice flown aboard the ISS using high throughput sequencing.

## **Materials and Methods**

### **Tissue Samples**

RNA samples were provided by the NASA Ames Research Center. RNA was extracted from the spleen and liver of 35-week-old female C57BL/6Tac mice that were either housed in the ISS environmental simulator (ground control, n=5), or flown aboard the ISS via SpaceX-4 (n=5). Tissues from flight animals were collected on board the ISS 21-22 days post-launch in flight animals while tissues from ground control animals were processed similarly on a four-day delay. Upon collection, spleens and livers were stored at -4°C in RNAlater (LifeTechnologies, Carlsbad, CA) for at least 24 hours and then stored at -80°C. RNA extraction was performed according to manufacturer's instructions with the RNeasy mini column (QIAGEN, Hilden, Germany) and stored at -80°C. Animal care and experimental procedures were approved by the Institutional Animal Care and Use Committee at the NASA Ames Research Center.

### **Illumina MiSeq Sequencing**

RNA samples were subjected to Illumina MiSeq sequencing at the Kansas State University Integrated Genomics Facility. Ig-specific primer amplification was not utilized. Illumina MiSeq with paired reads of 300 base pairs was performed on size selected (275-800 nt) total RNA isolated from the liver and spleen of three ground control and three flight animals based on highest RIN



values (Ground animals: G1, G2, G3; Flight animals: F1, F2, F3). Illumina MiSeq data from both spleen and liver are available by NASA GeneLab (<https://genelab.nasa.gov>, GLDS-ID Pending).

### **Bioinformatic Workflow**

Illumina MiSeq sequencing reads were processed as described previously (56). Briefly, FASTQ files were imported into CLC Genomics Workbench v9.5.1 (<https://www.qiagenbioinformatics.com/>) and were quality trimmed and filtered to remove sequences less than 40 nt in length. Paired-end sequences and overlapping-paired (merged) sequences were mapped to both V-gene segment references obtained from the ImMunoGeneTics (IMGT) database (251 IgH segments, 164 Ig $\kappa$  segments), and to entire IgH and Ig $\kappa$  loci obtained from NCBI (NC\_000078.6, 113258768 to 116009954, and NC\_000072.6, 67555636 to 70726754, respectively). Mapped sequencing reads were submitted to the IMGT HighV-Quest tool for characterization of functionality and junctional analysis. A motif search was performed in CLC on IgH sequences that were identified by IMGT as productive to determine their respective constant regions.

### **Gene Segment Usage**

Sequencing reads were analyzed using the IMGT HighV-Quest tool (78). V-gene segment usage was characterized as either productive or unknown functionality, where a read was considered productive if it was in frame and did not contain a premature stop codon as defined by IMGT. Sequences that did not fit the C-xx-W motif in non-class switched H-CDR3 or C-xx-F in  $\kappa$ -CDR3 were assigned an unknown functionality. D- and J-gene segment, and constant region usage was assessed in productive reads only. Reads assigned to multiple C57BL/6 V-gene

segments were tabulated using a weighted distribution. Reads containing only one possible V-gene segment were assigned a count of one. Reads containing two possible V-gene segments were assigned a count of 0.5 for each potential V-gene segment. Reads containing more than two potential V-gene segments were excluded from V-gene analysis. Reads assigned to a single non-C57BL/6 D/J-gene segments or multiple C57BL/6 D/J-gene segments were reclassified as undetermined and kept for analysis. J-gene segments in which less than six nucleotides were identified were also classified as undetermined. Percent of repertoire was determined for each individual animal by dividing the number of sequencing reads for each gene segment by the total number of sequencing reads mapped to all gene segments.

### **Gene Segment Combination & CDR3 Analysis**

Reads assigned to non-C57BL/6 V-gene segments or multiple C57BL/6 V-gene segments were removed from our V(D)J combination analyses. V(D)J combination analyses were performed on productive sequencing reads and visualized through the use of bubble charts (Microsoft Excel) and/or circos graphs from Circos Online (79). Percent repertoire was used to detail individual bubble charts and the average of percent repertoire was used when combining mice from each treatment group for V(D)J bubble chart analysis. CDR3 amino acid sequence was presented as percent of repertoire as described above and by a highest to lowest ranking of abundance.

CDR3 nucleotide alignments were created using the MAFFT multiple sequence alignment program (80). A V-D-J-gene segment combination was selected from the top ten percent most-represented gene segments across all individuals for both heavy and light chain. From each individual group, unique nucleotide sequences were isolated and aligned to their respective

germline sequences provided by IMGT. Individual alignments were then stacked within their treatment groups and germline gaps were adjusted for consistency across treatment groups.

### **Complementarity Determining and Framework Region Mutation Analysis**

Nucleotide substitution mutation data for complementarity determining and framework regions for IgH and Igk were obtained from the IMGT HighV-Quest tool. Any mutations involving degenerate bases were removed. Nucleotide range (in base pairs), number of reads containing at least one mutation, total number of substitution mutations, and number of mutations per base pair position were determined for each region. Comparative values for each combination of region, Ig location, and treatment group were determined by calculating the average of values contained in each combination's respective replicates (n=3).

### **Statistical Analysis and Representation of Data**

Differences in individual gene segment usage were assessed by Student's *t*-test in SAS University Edition ([www.sas.com](http://www.sas.com)). Pairwise linear regressions, normality tests, and two-way ANOVA were performed in GraphPad (version 6.0). Dot plots and bar graphs were generated in GraphPad using mean values and standard deviation. Heat maps of gene segment usage were generated in Microsoft Excel.

## **Results**

### **V-Gene Segment Usage**

B cells originate in the bone marrow from hematopoietic precursors, traffic through the periphery and enter the spleen where they are further selected and mature (81). To view a snapshot

of the impact that spaceflight has on the splenic Ig repertoire of unimmunized mice, we sequenced total splenic RNA isolated from three ground control animals and three animals flown aboard the ISS. We assessed the composition of individual IgH and Igκ sequences. In spleen, between 104,135 and 149,675 IgH, and between 105,374 and 172,660 Igκ sequencing reads of productive or unknown functionality were detected in ground animals, while between 66,909 and 181,703 IgH, and between 82,653 and 108,889 Igκ were detected in flight animals (Table 4.1).

The V-gene segment contributes to the combinatorial diversity of the Ig repertoire in part due to the large number of possible V-gene segments that could be selected within an individual B cell. Among the six study animals 127 VH- and 100 Vκ-gene segments were detected. Overall, the frequency of highly abundant V-gene segments and less frequently identified V-gene segments were similar between treatment groups (Appendix A.6). Despite a general similarity, a pairwise comparison of animals within treatment groups showed low-to-moderate levels of VH-gene segment correlation (Ground  $R^2$ : 0.3378-0.6862, p-values:<0.001; Flight  $R^2$ : 0.1100-0.3673, p-values:0.0001-<0.0001) that demonstrates that there is animal-to-animal variation (Table 4.2). A stronger correlation was seen in Vκ-gene segments (Ground  $R^2$ : 0.661-0.738, p-values <0.0001; Flight  $R^2$ : 0.466-0.603, p-values <0.0001) (Table 4.2). When comparing the average abundance of V-gene segments from ground and flight animals an  $R^2$  of 0.5833 was observed in VH ( $p$  =<0.0001) and 0.830 was observed in Vκ ( $p$  =<0.0001) (Table 4.2).

When comparing VH-gene usage among all animals, nine gene segments represented over five percent of the repertoire in at least one animal (Figure 4.1A). No one gene segment was found at over the five percent level in all six animals. V1-53 was found in over five percent in five animals, V9-3 was over five percent in four animals, V1-26 and V3-6 in two animals, and the remaining (V1-78, V6-3, V5-4, V1-15, V1-19) were found at high levels in only one animal.

Within  $V_{\kappa}$ , 9 gene segments represented over five percent of the repertoire in at least one animal (Figure 4.1B). The most abundant gene segment, V5-39 comprised over 15 percent of  $V_{\kappa}$  usage in all six animals. No other  $V_{\kappa}$  represented over five percent of the repertoire in all six animals. Three gene segments (V6-25, V4-61, V6-13) were found at five percent abundance in one animal and less than one percent abundance in all other animals. There was no statistical difference in top VH- or  $V_{\kappa}$ -gene segment usage between ground and flight animals (Student's *t*-test,  $p=0.0656-0.8280$ ).

We also attempted to assess the antibody repertoire in the liver because of its role in fetal B-cell development. Only 471-1,287 Ig $\kappa$  sequencing reads were detected in ground control animals and 309-376 Ig $\kappa$  sequencing reads were detected in flight animals (Table 4.3). We did not characterize the heavy chain in the liver due to the low number of IgH sequencing reads that we detected. We assessed  $V_{\kappa}$ -gene segments that represented over five percent of the repertoire in the spleen or liver and found 15 gene segments. Only four of those gene segments were shared between tissues (Figure 4.2). Overall, the average usage of these top  $V_{\kappa}$ -gene segments showed modest to high correlation between liver and spleen in ground animals ( $R^2= 0.3167$ ,  $p=0.0290$ ) and flight animals ( $R^2= 0.7994$ ,  $p<0.0001$ ). Analysis of statistical differences between individual  $V_{\kappa}$ -gene segments representing over five percent of the repertoire was not undertaken due to low read counts in the liver datasets.

### **D- and J-Gene Segment & Constant Region Usage**

Heavy chain Ig diversity is also achieved by using, modifying and splicing of D- and J-gene segments. To determine if space flight affected these processes we also assessed D- and J-gene usage. The most commonly detected D-gene segment in both ground control and flight animals was D1-1, comprising between 30.6% to 46.8% of the repertoire (Figure 4.3A, Appendix

A.7A). D2-4, D2-3, D4-1, and D2-5 were detected at similar levels among ground control and flight animals between 3.56% and 11.39% of the repertoire. D3-1, D3-2, D6-2, D5-1, and D5-5 were detected the least often with levels between 3.6% and 0% of the repertoire. Because of extensive modification of D-gene segments during IgH rearrangement, D-gene segments were unable to be determined for between 24.4% and 36.6% of the repertoire for all animals. There was no statistical difference in D-gene segment usage between ground and flight animals (Student's *t*-test,  $p=0.1542-0.9840$ ). D-gene segment usage was highly correlated between ground and flight animals (linear regression,  $R^2= 0.9939$ ,  $p<0.0001$ ).

Additional Ig variability is gained from the inclusion of different J-gene segments. Within IgH, the distribution of J-gene segment usage was less uniform than D-gene segment usage in both ground and flight animals. There was no consensus on the most abundantly expressed gene segment as each of the four JH-gene segments was the most abundant segment in at least one ground or flight animal (Figure 4.3B, Appendix A.7B). While there was no consensus use of a particular J-gene segment, usage of any J-gene segment was between 13% and 40%, showing that usage is relatively uniform. Student's *t*-tests were performed to determine whether differences in individual gene segment usage were significant. Significant differences were only found between J2 and J4 in ground animals (Student's *t*-test,  $p=0.0267$ ), and J2 and J3 in ground animals (Student's *t*-test,  $p=0.0429$ ). As the most abundant J $\kappa$ -gene segment, J5, comprises between 31.9% and 48.1% of the repertoire, and is the most abundant gene segment in five out of six study animals (Figure 4.3C, Appendix A.7C). Interestingly, J1 ranked second most abundant in all ground animals while it ranked third in all flight animals. There were no statistical differences in individual JH- or J $\kappa$ -gene segment usage between ground control and flight animals (Student's *t*-test,

p=0.1001-0.9239). Linear regression revealed correlation between JH usage of ground and flight animals for Ig $\kappa$  ( $R^2=0.9202$ ,  $p=0.0098$ ), but not IgH ( $R^2=0.7658$ ,  $p=0.0520$ ).

Ig isotype composition can provide insight into the developmental stage of B cells. Because animals in this experiment were specific-pathogen free, it is unsurprising that IgM predominated with between 62.38% and 82.81% of the repertoire (Figure 4.3D, Appendix A.7D). IgG was the second most prominent isotype which trended towards a higher percentage of the repertoire in flight animals although the difference was not statistically significant (Student's *t*-test,  $p=0.2150$ ). Except for a relatively high expression of IgA in flight mouse two, IgA was detected in between 1.43% and 4.58% of the repertoire, IgD and IgE were detected less than one percent in all animals. There were no statistical differences in Ig isotype frequency between ground and flight animals (Student's *t*-test,  $p=0.1075-0.8277$ ). There was a high correlation between ground and flight constant region usage (linear regression,  $R^2=0.9734$ ,  $p=0.0019$ ).

## **V(D)J Combinations**

V(D)J family combinations were examined as another way to determine if recombination of Ig gene segments was affected by spaceflight. We visualized V(D)J-gene segment combinations using both bubble charts and circos plots for both IgH and Ig $\kappa$ . IgH showed more variation between ground control and flight animals compared to Ig $\kappa$  when looking at the most common gene family combinations. Circos plots allowed us to examine top V-gene families, J-gene segments, and V/J pairing frequency in both IgH and Ig $\kappa$ .

For ease of display, we first grouped together all V-gene segments into their respective family. D- and J-gene segments remained as individuals. V1 was the most common IgH gene family used in all mice comprising  $\geq 42\%$  of the V-gene family use in ground animals and  $\geq 50\%$

of the repertoire in flight animals (Figure 4.4A-B, Appendix A.8A-F). In ground-treatment animals, V9 was the second most common gene family in ground mouse one and ground mouse two, while V2 was the second most common gene family in ground mouse three (Figure 4.4A, Appendix A.8A-C). In flight animals, the second most common V-gene family used was unique among the three animals (flight one: V3, flight two: V5, and flight three: V9) (Figure 4.3B, Appendix A.8D-F). The third most common V-gene family was also unique among the ground treatment animals being V2, V3, and V9 for ground mouse one, ground mouse two, and ground mouse three respectively (Figure 4.4A, Appendix A.8A-C). V2 was the third most common family in flight mouse two and flight mouse three, while V5 was the third most common gene family used for flight mouse one (Figure 4.4B, Appendix A.8D-F).

We found that the most common V/D/J combinations were correlated with the most frequently used gene families or segments within the repertoire. When averaging among repertoires, the most common IgH combination in ground-treatment animals was V9/D1/J1 (6.86%), though this combination was only in the top five most frequent combinations in ground mouse one and ground mouse two. The most common average combination in flight-treatment animals was V1/D1/J4 (8.27%), though this combination was only detected in the five most common combinations for one animal (Figure 4.4A-B, Appendix A.9A-B). The V1/D1/J2 combination was shared among the top five gene family combinations in all mice; representing 5.54% of the repertoire in ground-treatment animals and 7.68% in the flight-treatment animals (Appendix A.9A-B). A notable difference between ground and flight treatment groups was the usage of the V9-gene family. This family represented the top average VH-gene family used in the ground-treatment animals as well as the top combination in ground mouse one and ground mouse



two (Figure 4.4A, Appendix A.9A), but only appeared once as the top gene family combination used for flight-treatment animals (Figure 4.4B, Appendix A.9B).

We also examined the top five V/J pairing frequencies for IgH (Figure 4.4A-B, Appendix A.8A-F, Appendix A.9C-D). For ground animals, there were six unique pairings represented. The V1 family was used for four of the six unique V/J pairings compiled. V9 and V2 were also used. Of the six unique pairings, four were shared among all three mice (V1/J1, V1/J2, V1/J3, and V1/J4). One was shared among two mice (V9/J1), and one (V2/J3) was found only in ground mouse three's five most common combinations (Figure 4.4A, Appendix A.8A-C, Appendix A.9C). For flight animals, seven unique pairings were found in the five most common pairs. V1 was again the overwhelmingly most common V family with four of the seven unique pairings including it. V3, V2, and V9 were all used by a single animal. Of the seven unique pairs, four were shared among all three mice (V1/J1, V1/J2, V1/J3, and V1/J4). These pairings were also among the most common in the ground-treatment group. Three (V3/J3, V2/J4, V9/J1) were found in a single mouse's five most common pairings (Figure 4.4B, Appendix A.8D-F, Appendix A.9D).

We undertook similar analysis for Ig $\kappa$  sequences (Figure 4.4C-D, Appendix A.8G-L, Appendix A.9E-F). The most common V $\kappa$ -gene family expressed in all mice was V5, representing over one-fifth of the repertoire. The second most common V $\kappa$ -gene family in ground animals was V3 while in flight animals it was (Figure 4.4C-D, Appendix A.8G-L, Appendix A.9E-F). The third most represented V $\kappa$ -gene families were V6 for ground mouse one and V4 for ground mouse two and three (Figure 4.4C, Appendix A.8G-I). In flight animals, the third most common V $\kappa$ -gene family was V6 for flight mouse one and flight mouse two, and V3 for flight mouse three (Figure 4.4D, Appendix A.8J-L).

Unlike IgH, Igκ expressed more variety in V/J pairing variety. In Igκ there were four pairings in ground animals (V3/J5, V5/J1, V3/J2, V2/J4) that were not shared with the top five flight-animal pairings (Figure 4.4C-D, Appendix A.8G-L, Appendix A.9E-F). There were also five pairings (V1/J2, V6/J5, V4/J2, V6/J1, V4/J4) found in flight animals not found in the top five ground-animal pairings (Figure 4.4C-D, Appendix A.8G-L, Appendix A.9E-F).

The correlation of the average percent of repertoire for V/J combinations between ground and flight animals was higher in Igκ, with a  $R^2$  of 0.8904 ( $p < 0.0001$ ), whereas IgH had a  $R^2$  of 0.3229 ( $p < 0.0001$ ) (linear regression). V/J combination usage among animals within each treatment group also showed stronger correlation within Igκ than IgH in both ground and flight treatment groups (Table 4.4).

### **CDR3**

CDR3 is important for conferring diversity in Ig specificity. Therefore, we assessed whether spaceflight had an impact on several properties of CDR3 because changes in CDR3 could affect host ability to respond to antigen. We found that CDR3 length in IgH was highly varied. The length ranged from one amino acid to 35 (Figure 5A). The average CDR3 lengths for all the ground control animals was  $12 \pm 0$ . The average CDR3 lengths for flight animals was also  $12 \pm 1$ . The CDR3 lengths were not normally distributed in the flight or ground animals (flight  $p = < 0.0001$ , ground  $p = 0.0128$ ) with the majority of CDR3 lengths falling between 11 and 14 AAs. We also examined the heavy-chain CDR3 length by isotype (Table 5) and by treatment group. We found no significant difference by treatment group or by isotype (two-way ANOVA, interaction  $p = 0.8141$ , isotype  $p = 0.4589$ , treatment  $p = 0.6225$ ).

Kappa-chain CDR3 length was conserved at nine amino acids with 90.8 to 94.4 percent of all light chains having CDR3 sequences that were nine amino acids in length (Figure 5B). Only a small percentage of light chains had eight amino acids (2.7 to 6.7%), or ten amino acid long (0.9-2.0%) CDR3s. Ground mouse one (0.9%), was an exception and was enriched for 11 amino acid CDR3s (2.3%) compared to other animals (0.2-0.7%). CDR3 lengths over 20 amino acids were not displayed (25, 27, 29, 30, 32, 33, 35, 36, 37, 39). CDR3 sequences of these lengths were often only detected in one animal and expression of these lengths did not exceed 0.01% of the repertoire. Additionally, these sequences may have been identified in error as many of these sequences contain intervening phenylalanine residues within the conserved kappa-chain CDR3 C-XX-F motif.

There was little overlap among IgH CDR3s regardless of treatment (Figure 6A, C). Of the top five CDR3s found in each animal, we identified 26 unique CDR3 AA sequences. Of those 26 CDR3s, four were found in all six animals. Two additional sequences were identified in three animals. One of those sequences, CASHGSSYLAWFAYW, was found in only flight animals and not found in any ground animals. Six CDR3s were found in two animals, and the remainder were found in a single animal. The vast majority of CDR3 sequences detected were unique to each animal (6,661-9,270 sequences), though there was a small amount of overlap among animals (103-163 sequences). For flight animals, 78 sequences were found in all three animals and 70 were found in all three ground animals. Of the sequences found in all three animals per treatment group, only 20 sequences were detected in all six animals.

There was considerable overlap in the top five Igκ CDR3 amino acid sequences of each animal (Figure 6B, D). Seventeen total top CDR3 sequences were identified and all CDR3 sequences were identified in every animal. One CDR3 was the most abundant sequence in five

animals (CQNGHSFPLTF), still ranking third in the remaining animal (F1). Ground control and flight animals shared six and five sequences that ranked within the top 20 CDR3 in all animals, respectively. Overall, between 1,515-2,733 unique CDR3 were detected in ground control animals, and between 1,377-1,756 unique CDR3 were detected in all flight animals (Figure 6). Of the 685 and 556 CDR3 shared in all three ground control and flight animals, respectively, 427 were shared between all six ground and flight animals.

A CDR3 nucleotide sequence alignment of one of the top V-D-J-gene segment combinations demonstrates significant variability among mice such that any variability between ground and flight treatment groups cannot be determined with confidence (Figure 7). Additional data sets are needed in order to assess the effect of spaceflight on CDR3 formation.

### ***3.5 Mutations in Complementarity Determining and Framework Regions***

We also examined mutation frequencies in CDR and framework (FW) regions for each animal because mutations can affect Ig specificity. Mutation frequencies were normalized by animal and Ig region (FW1-3, CDR1-3) by dividing the percent of total substitution mutations (total mutations/total productive reads) by respective region length (Figure 8). There were no significant differences between the mutation frequencies of ground control and flight animals for any of the Ig regions in both IgH (Student's *t*-test,  $0.2317 < p < 0.8516$ ) and Igκ (Student's *t*-test,  $0.4562 < p < 0.6390$ ). When comparing the substitution mutation frequency across Ig regions, more substitution mutations occurred in CDR3 compared to other regions in IgH (Student's *t*-test; all  $p < 0.0001$ ) and Igκ (Student's *t*-test; all  $p < 0.0001$ ).

### CDR3

CDR3 is important for conferring diversity in Ig specificity. Therefore, we assessed whether spaceflight had an impact on several properties of CDR3 because changes in CDR3 could affect host ability to respond to antigen. We found that CDR3 length in IgH was highly varied. The length ranged from one amino acid to 35 (Figure 4.5A). The average CDR3 lengths for all the ground control animals was  $12 \pm 0$ . The average CDR3 lengths for flight animals was also  $12 \pm 1$ . The CDR3 lengths were not normally distributed in the flight or ground animals (flight  $p < 0.0001$ , ground  $p = 0.0128$ ) with the majority of CDR3 lengths falling between 11 and 14 AAs. We also examined the heavy-chain CDR3 length by isotype (Table 4.5) and by treatment group. We found no significant difference by treatment group or by isotype (two-way ANOVA, interaction  $p = 0.8141$ , isotype  $p = 0.4589$ , treatment  $p = 0.6225$ ).

Kappa-chain CDR3 length was conserved at nine amino acids with 90.8 to 94.4 percent of all light chains having CDR3 sequences that were nine amino acids in length (Figure 4.5B). Only a small percentage of light chains had eight amino acids (2.7 to 6.7%), or ten amino acid long (0.9-2.0%) CDR3s. Ground mouse one (0.9%), was an exception and was enriched for 11 amino acid CDR3s (2.3%) compared to other animals (0.2-0.7%). CDR3 lengths over 20 amino acids were not displayed (25, 27, 29, 30, 32, 33, 35, 36, 37, 39). CDR3 sequences of these lengths were often only detected in one animal and expression of these lengths did not exceed 0.01% of the repertoire. Additionally, these sequences may have been identified in error as many of these sequences contain intervening phenylalanine residues within the conserved kappa-chain CDR3 C-XX-F motif.

There was little overlap among IgH CDR3s regardless of treatment (Figure 4.6A, C). Of the top five CDR3s found in each animal, we identified 26 unique CDR3 AA sequences. Of those

26 CDR3s, four were found in all six animals. Two additional sequences were identified in three animals. One of those sequences, CASHGSSYLAWFAYW, was found in only flight animals and not found in any ground animals. Six CDR3s were found in two animals, and the remainder were found in a single animal. The vast majority of CDR3 sequences detected were unique to each animal (6,661-9,270 sequences), though there was a small amount of overlap among animals (103-163 sequences). For flight animals, 78 sequences were found in all three animals and 70 were found in all three ground animals. Of the sequences found in all three animals per treatment group, only 20 sequences were detected in all six animals.

There was considerable overlap in the top five Igk CDR3 amino acid sequences of each animal (Figure 4.6B, D). Seventeen total top CDR3 sequences were identified and all CDR3 sequences were identified in every animal. One CDR3 was the most abundant sequence in five animals (CQNGHSFPLTF), still ranking third in the remaining animal (F1). Ground control and flight animals shared six and five sequences that ranked within the top 20 CDR3 in all animals, respectively. Overall, between 1,515-2,733 unique CDR3 were detected in ground control animals, and between 1,377-1,756 unique CDR3 were detected in all flight animals (Figure 4.6C). Of the 685 and 556 CDR3 shared in all three ground control and flight animals, respectively, 427 were shared between all six ground and flight animals.

A CDR3 nucleotide sequence alignment of one of the top V-D-J-gene segment combinations demonstrates significant variability among mice such that any variability between ground and flight treatment groups cannot be determined with confidence (Figure 4.7). Additional data sets are needed in order to assess the effect of spaceflight on CDR3 formation.

## **Mutations in Complementarity Determining and Framework Regions**

We also examined mutation frequencies in CDR and framework (FW) regions for each animal because mutations can affect Ig specificity. Mutation frequencies were normalized by animal and Ig region (FW1-3, CDR1-3) by dividing the percent of total substitution mutations (total mutations/total productive reads) by respective region length (Figure 4.8). There were no significant differences between the mutation frequencies of ground control and flight animals for any of the Ig regions in both IgH (Student's *t*-test,  $0.2317 < p < 0.8516$ ) and Igk (Student's *t*-test,  $0.4562 < p < 0.6390$ ). When comparing the substitution mutation frequency across Ig regions, more substitution mutations occurred in CDR3 compared to other regions in IgH (Student's *t*-test; all  $p < 0.0001$ ) and Igk (Student's *t*-test; all  $p < 0.0001$ ).

## **Discussion**

Spaceflight and ground-based analog models induce phenotypic and functional changes in T- and B- lymphocyte populations. Spaceflight also affects bone marrow, the site of B-cell differentiation and development. Therefore, we wanted to know whether the stress and physiological changes associated with spaceflight would affect the normal development of the highly-diversified and highly-specific antigen receptors on B-lymphocytes. If so, the ability to respond to pathogens might be affected. We characterized the antibody repertoire of C57BL/6Tac mice flown aboard the ISS and ground control animals using HTS and RNA-Seq.

HTS studies of antibody repertoires typically employ polymerase chain reaction amplification of Ig specific sequences from sorted B-cell populations. We assessed the B cell repertoire in whole spleen tissue because of the limitations of the primary science, a verification flight of mouse housing hardware. This precluded the sorting of cell populations. We previously

showed that similar data could be generated using whole spleen tissue compared to whole spleen cell suspensions (56). Although we do not account for B-cell subpopulations (82, 83), we do measure the total splenic Ig repertoire. Additionally, since we did not use specific amplification of Ig sequences the depth of sequencing was not as high as some have accomplished looking at Ig gene usage (82-84) Nevertheless, comparisons of amplified and unamplified data sets by our lab show reasonable correlations of the data (In preparation). Amplification with multiplex Ig specific primers may introduce amplification bias as primers may bind with varying efficiency to V-gene segments, although there have been recent advances in experimental approach to address amplification bias (57). We also found that the type of RNA-Seq analysis we are using in the assessment of younger C57Bl/6J mouse Ig gene usage, correlated well with that of studies using amplification (56). Therefore, we feel we have a reasonable snapshot of B-cell receptors present in the spleens of 35-week-old female mice. In addition, the sample preparation allows additional data mining of valuable mouse samples.

In both humans and mice, early B-cell development occurs in the fetal liver prior to postnatal development in the bone marrow. We attempted to determine the Ig repertoire within the adult liver of the ground and flight animals in comparison to the splenic Ig repertoire. Unfortunately, few Ig sequences were recovered in liver samples suggesting few B cells are actually resident or circulating in the liver under normal, steady-state conditions. The liver kappa chain V-gene repertoire did correlate some with the usage in the spleen and probably reflects circulating B-cell Ig expression but we did not confirm that. We focused our efforts on the spleen data.

While previous analyses by our lab used splenic mRNA pooled from four animals (Rettig et al., Submitted for Publication in PloS one, (56)), the current study assessed individual mice and



exhibited significantly more mouse-to-mouse variation than one might expect in inbred mice; even within treatment groups and compared to pooled mouse samples. Overall, V-gene segment usage correlated when analyzed using pairwise linear regression of animals within ground and flight treatment groups and there did not appear to be an impact of spaceflight on B-gene segment use. It is possible that differences in gene segment usage in ground and flight animals would be observed upon immunization. Studies on the effects of spaceflight in an immunized amphibian model showed altered VH-gene family and Ig class usage (74, 75).

We performed a number of analyses to determine if the Ig gene rearrangement process was affected by spaceflight. Averaged V-gene segment usage between the ground control and flight animals was moderately to highly correlated (VH:  $R^2=0.5833$ ,  $p<0.0001$ ; V $\kappa$ :  $R^2=0.830$ ,  $p<0.0001$ ). Many of the most abundant V-gene sequences were shared in flight and ground animals and there was no statistically significant difference in usage of individual top V-gene segments between ground control and flight animals (Student's *t*-test,  $p=0.0656-0.8280$ ). Similarly, no differences in D-, JH- or J $\kappa$ -gene segment usage and IgH constant region usage was seen between ground and flight animals.

We also examined whether spaceflight would affect the V/J-gene segment combinations that normally occur in specific pathogen-free mice. These too, were not different between ground control and flight animals for IgH and Ig $\kappa$ . V/J-gene segment combinations were moderately to highly correlated (IgH:  $R^2=0.3229$ ,  $p<0.0001$ ; Ig $\kappa$ :  $R^2=0.8904$ ,  $p<0.0001$ ). Nevertheless, we did note that while the most abundant VH-gene family in ground animals is VH9, this gene family did not appear in the top gene family combinations in flight animals. We will need a larger sample size to determine if this is a spaceflight effect or if it just reflects the large mouse-to-mouse variation we observe. An assessment of the impact of hypergravity on the similarly assembled T

cell receptor repertoire of neonatal mice showed low correlation of individual Beta chain V- and J-gene segment recombination frequencies between control animals and animals subjected to centrifugation. Eighty-five percent of gene V/J-gene segment combinations were not shared among the two treatment groups (85) and the differences could be attributed to changes in somatic recombination machinery under altered gravity conditions (85, 86).

The combinatorial diversity of Ig was shown through the assessment of overall CDR3 sequence overlap among animals, as a large number of sequences were unique to only one animal. Little overlap was observed in the top 5 H-CDR3 sequences within all six ground and flight animals which totaled to 26 CDR3 sequences. Thirteen of these CDR3 sequences were only identified in one animal. More overlap was observed in the top 5  $\kappa$ -CDR3 sequences within all six ground and flight animals which totaled 17 CDR3 sequences, which were identified in all animals.

The mice studied in this investigation were not challenged and were housed under specific-pathogen free conditions. Mutational frequency in Ig is normally associated with antigen stimulation (87). Therefore, we did not expect that these mice were undergoing high amounts of somatic mutations. The majority of the mutations detected occurred in CDR3 in both ground and flight mice. It is possible that we will see differences in mutation frequency between ground control and flight animals after experimental immune challenge. The frequency of somatic hypermutations in *P. waltl* immunized in space was slightly lower than animals immunized on earth (76). An experiment with antigen challenge of the Ig repertoire will be necessary to test this hypothesis.

The animals used in this experiment were older (35 weeks). The expression of genes necessary for Ig recombination do go down as mice age (88). Therefore, it may be possible that we do not see differences between the treatment groups because the perturbations of spaceflight are not enough to disrupt the reduced B cell differentiations occurring in 35-week-old mice under

normal steady-state conditions. Additional experiments, especially with young mice, will be needed to test this hypothesis.

Animals in this experiment were older and supplied from a different vendor (Taconic) than the 9-11-week-old C57BL/6J mice from Jackson Laboratories used in our previous experiments ((56), (Rettig et al., Submitted for Publication in PloS one)). To assess whether Ig repertoire differences existed between the older C57BL/6Tac mice and the younger C57BL/6J mice, we compared gene segment usage between the two mouse cohorts. Because no differences in top V-gene segments, (D)J-gene segments, and constant region usage were detected between RR1 ground control and flight animals, all six animals were pooled and compared to the C57BL/6J cohort. We selected the 25 most abundant V-gene segments from both RR1 and C57BL/6J cohorts, resulting in 35 top IgHV and 32 top IgKV. We found that six of 35 IgHV and nine of 32 IGKV were expressed at significantly different levels within the repertoire between the two cohorts (Student's *t*-test). These differences are largely driven by gene segments that are highly expressed in the RR1 cohort such as IGHV1-53, which represented  $11.53 \pm 6.48\%$  of the repertoire in RR1 animals and only  $3.09 \pm 0.65\%$  of the repertoire in the C57BL/6J cohort (Student's *t*-test,  $p=0.0237$ ). This is even more pronounced in IGKV5-39, which represented  $21.59 \pm 6.63\%$  of the repertoire in RR1 animals and only  $5.39 \pm 6.63\%$  of the repertoire in the C57BL/6J cohort (Student's *t*-test,  $p=0.0096$ ).

We also performed a linear regression of top V-gene segment usage, which showed poor correlation between cohorts in both IgH ( $R^2 = 0.1614$ ,  $p=0.0168$ ) and IgK ( $R^2 = 0.2348$ ,  $p=0.005$ ). D-gene segment usage was only significantly different IGHD1-1, which represented  $39.7 \pm 4.45\%$  of the repertoire in RR1 animals and  $24.5 \pm 4.45\%$  of the repertoire in the C57BL/6J cohort (Student's *t*-test,  $p=0.0040$ ). We found that J-gene segment usage varied between the two cohorts in three out of five IGKJ and no differences in IGHV were detected (Student's *t*-test). IgH constant

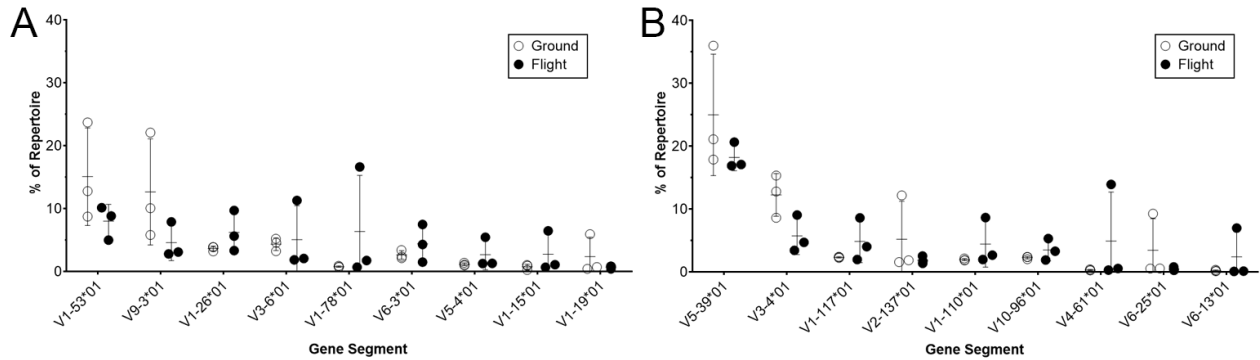
region usage was significantly different for IgD (RR1:  $0.63 \pm 0.22\%$ , C57BL6/J:  $4.60 \pm 0.69\%$  ; Student's *t*-test,  $p=0.0074$ ) and IgG (RR1:  $23.83 \pm 6.19\%$ , C57BL6/J:  $9.65 \pm 6.19\%$  ; Student's *t*-test,  $p=0.0074$ ).

Although we cannot determine whether these differences are attributed to differences in vendor or differences in age, it is likely that repertoire differences are driven by a more mature Ig repertoire within the RR1 animals, as a higher percentage of IgH sequences demonstrate class switching. Both cohorts were unimmunized and maintained under specific pathogen-free conditions.

In conclusion, we have been able to successfully characterize immunoglobulin gene segment usage and junctional diversity within the antibody repertoire of unimmunized C57BL/6Tac mice flown aboard the ISS. Individual gene segment usage remained similar among animals within and among treatment groups, with the most abundant gene segments being conserved across all animals. Gene segment combinations and CDR3 sequences were highly varied, demonstrating the combinatorial diversity of the antibody repertoire, but that variation reflects the dynamics of individualized selection of Ig molecules and not any impact of spaceflight. A larger sample size would help solidify this conclusion, but these data provide preliminary suggestions that the recombinatorial processes that lead to the diverse Ig repertoires in mice are not affected by a short trip to and stay on the ISS. These data do not preclude that differences in the Ig repertoires of ground and flight animals will not be seen during active immunization. Current studies in our lab aim to characterize antibody repertoire dynamics upon antigen challenge using a murine anti-orthostatic suspension model.

## Figures and Tables

**Figure 4.1 Expression of top V-gene segments**



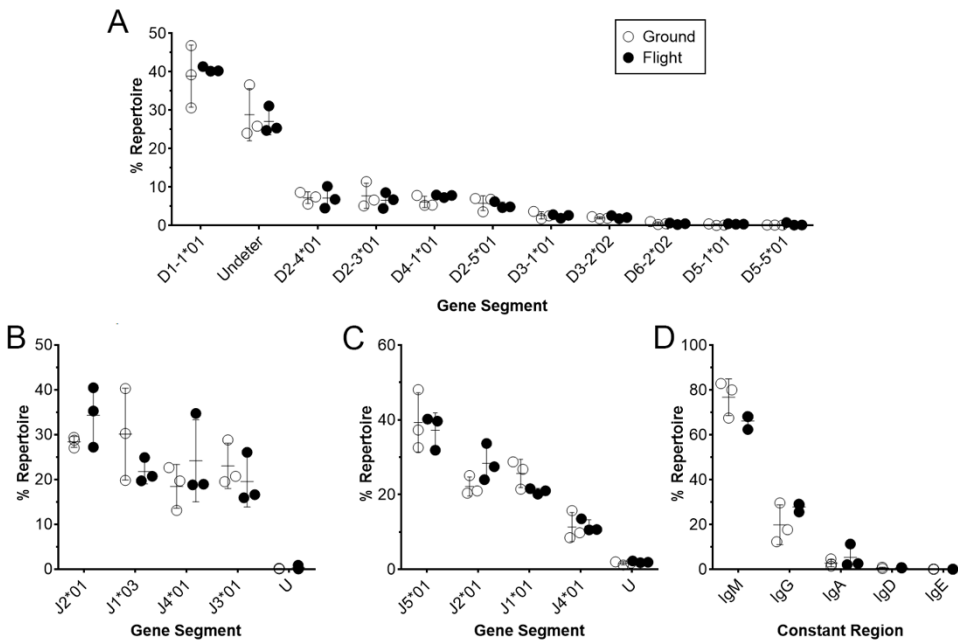
(A) VH- and (B) V $\kappa$ -gene segment usage for gene segments representing over five-percent of the repertoire in at least one animal within ground and flight treatment groups. No significant difference in individual gene segment usage was detected between ground control and flight treatment groups (Student's *t*-test,  $0.0656 < p\text{-value} < 0.8280$ ). Significant differences between gene segment usage of combined ground and flight animals were found. In IgH, V1-53 was more abundant than many of the top V-gene segments (V1-26, V3-6, V6-3, V5-4, V1-15, V1-19; Student's *t*-test, all  $p < 0.05$ ), and V1-26 was more abundant than V5-4, V1-15, and V1-19 (Student's *t*-test, all  $p < 0.05$ ). In Ig $\kappa$ , V5-39 was more abundant than all top V-gene segments (Student's *t*-test, all  $p < 0.05$ ), and V3-4 is more abundant than V1-117, V1-110, V10-96, V6-25 and V6-13 (Student's *t*-test, all  $p < 0.05$ ).

**Figure 4.2 Expression of top V $\kappa$ -gene segments from spleen and liver**

	LG1	LG2	LG3	LF1	LF2	LF3		SG1	SG2	SG3	SF1	SF2	SF3
V5-39*01	4	1	2	1	1	1	V5-39*01	1	1	1	1	1	1
V3-4*01	5	2	1	3	6	3	V3-4*01	2	2	2	5	5	2
V1-117*01	17	4	5	6	5	20	V1-117*01	6	7	7	2	3	10
V2-137*01	14	9	2	14	17	7	V2-137*01	16	9	3	8	13	19
V1-110*01	15	7	10	7	4	5	V1-110*01	9	10	11	14	8	3
V10-96*01	10	22	43	14	9	12	V10-96*01	8	6	10	4	6	12
V4-61*01	38	17	62	57	2	25	V4-61*01	55	49	55	54	2	37
V6-25*01	6	78	62	77	55	65	V6-25*01	3	38	35	29	57	54
V3-2*01	2	3	23	35	3	12	V3-2*01	19	8	21	20	7	13
V4-68*01	33	68	20	44	7	4	V4-68*01	40	33	4	37	21	4
V6-13*01	38	70	62	3	55	65	V6-13*01	51	65	67	3	74	75
V6-20*01	3	13	30	20	31	12	V6-20*01	12	35	29	15	29	23
V4-91*01	45	32	62	25	33	2	V4-91*01	45	41	31	33	18	7
V1-99*01	23	7	15	2	17	25	V1-99*01	50	42	66	66	76	61
V6-14*01	1	70	62	77	75	48	V6-14*01	49	69	57	69	55	66

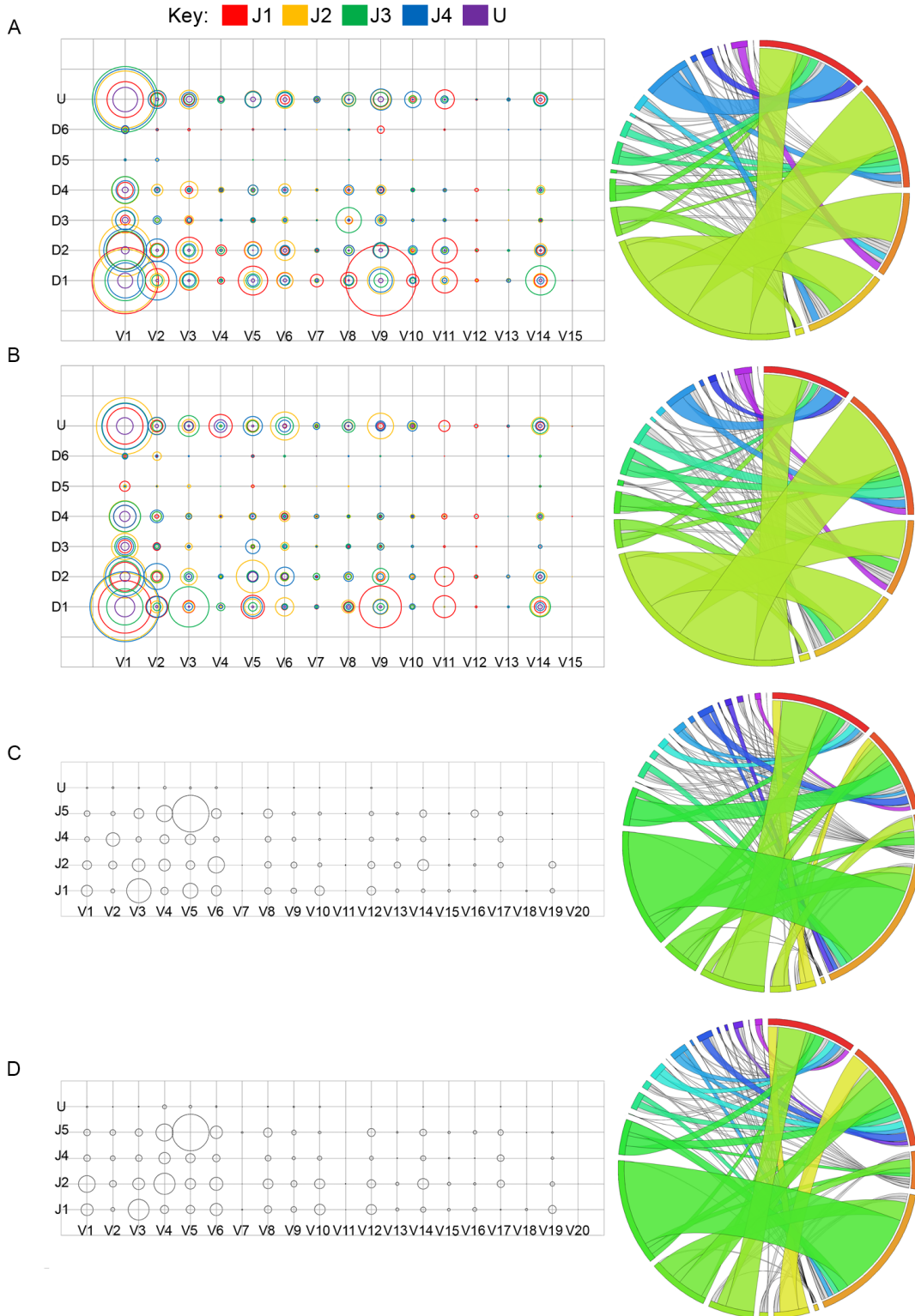
V $\kappa$ -gene segment usage for gene segments representing over five-percent of the repertoire in at least one animal from the liver or spleen of ground or flight animals are presented by rank. Liver ground (LG) and liver flight (LF) rankings are shown to the left and spleen ground (SG) and spleen flight (SF) rankings are shown to the right. V-gene segments are listed most frequent to least frequent. Dark red indicates higher rank moving to blue, lower percent rank.

**Figure 4.3 Expression of DH- and J-gene segments and IgH constant region usage**



(A) D-gene segment, (B) JH-gene segment, (C) Jκ-gene segment, and (D) IgH constant region usage in animals within ground and flight treatment groups are presented as percent of repertoire. No significant difference in individual gene segment usage was detected between ground control and flight treatment groups (Student's *t*-test, D: 0.1542<p-value<0.9840, JH: 0.2049<p-value<0.4782, IgH Constant Region: 0.1075<p-value<0.8277, Jκ: 0.1001<p-value<0.9239). Significant differences between D-gene segment usage of combined ground and flight animals were found between all gene segments except D2-4 with D2-3, D4-1, or D2-5, D-2-3 with D4-1 or D2-5, D4-1 with D2-5, and D3-1 with D3-2 (Students *t*-test, all significant p<0.05). Significant differences in J-gene segment usage of combined ground and flight animals were found. In IgH, J2 is more abundant than J3 and J4, both (Student's *t*-test, all p<0.05). In Igκ, significant differences in expression were found between all gene segments except between J1 and J2 (Student's *t*-test, all significant p<0.05). No significant differences in constant region usage were detected.

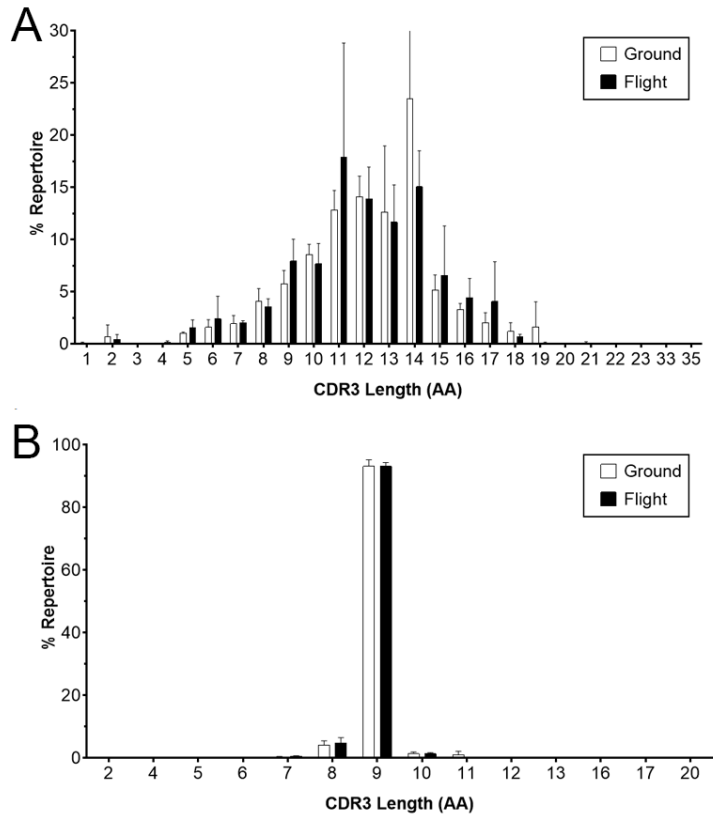
**Figure 4.4 Gene segment combinations in ground control and flight animals**





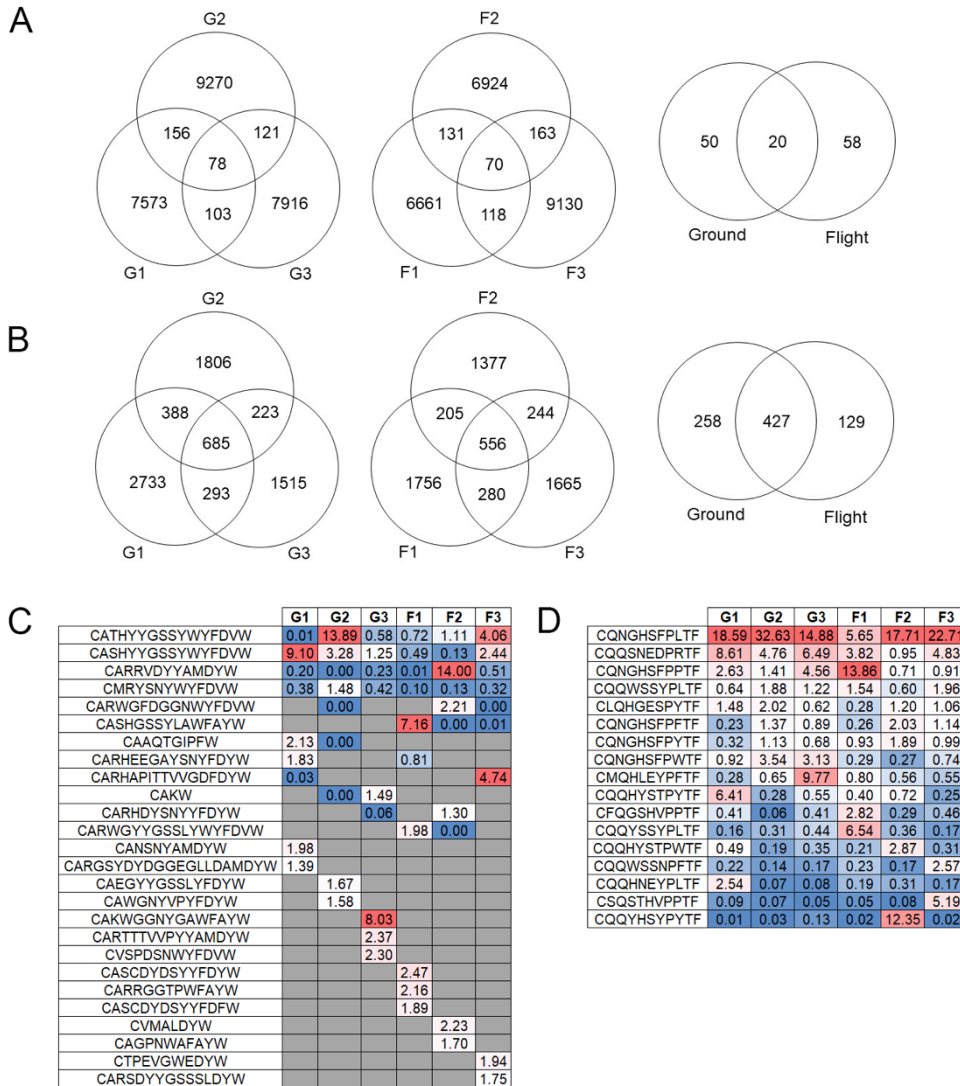
(A,B) Average IgH V/D/J combinations (bubble chart) and the V/J combinations (Circos plot) for ground treatment (A) animals and (B) flight animals. For bubble charts, V-gene family is represented along the x-axis, the D-gene segment is represented along the y-axis, and the J-gene segment is represented by a specific color. The size of the bubble corresponds to the average percent repertoire of the specific gene combination. Circos plots are read clockwise starting at the 12 o'clock position with J1 (red), J2, J3, J4, U, V1 (lime green), V2, V3, V4, V5, V6, V7, V8 (light blue), V9, V10, V11, V12, V13, V14, and V15 (sliver, no color). (C,D) Average Ig $\kappa$  V/J combinations for ground treatment (C) animals and (D) flight animals. For bubble charts, V-gene family is represented along the x-axis and J-gene segment is represented along the x axis. The size of the bubble corresponds to the average percent repertoire of the specific gene combination. Circos plots are read clockwise starting at the 12 o'clock position with J1 (red), J2, J4, J5, U, V1 (yellow), V2, V3, V4, V5, V6, V7 (sliver, no color), V8, V9, V10, V11, V12, V13, V14, V15, V16, V17, V18, V19 (light purple).

**Figure 4.5 CDR3 length in IgH and Igk sequences**



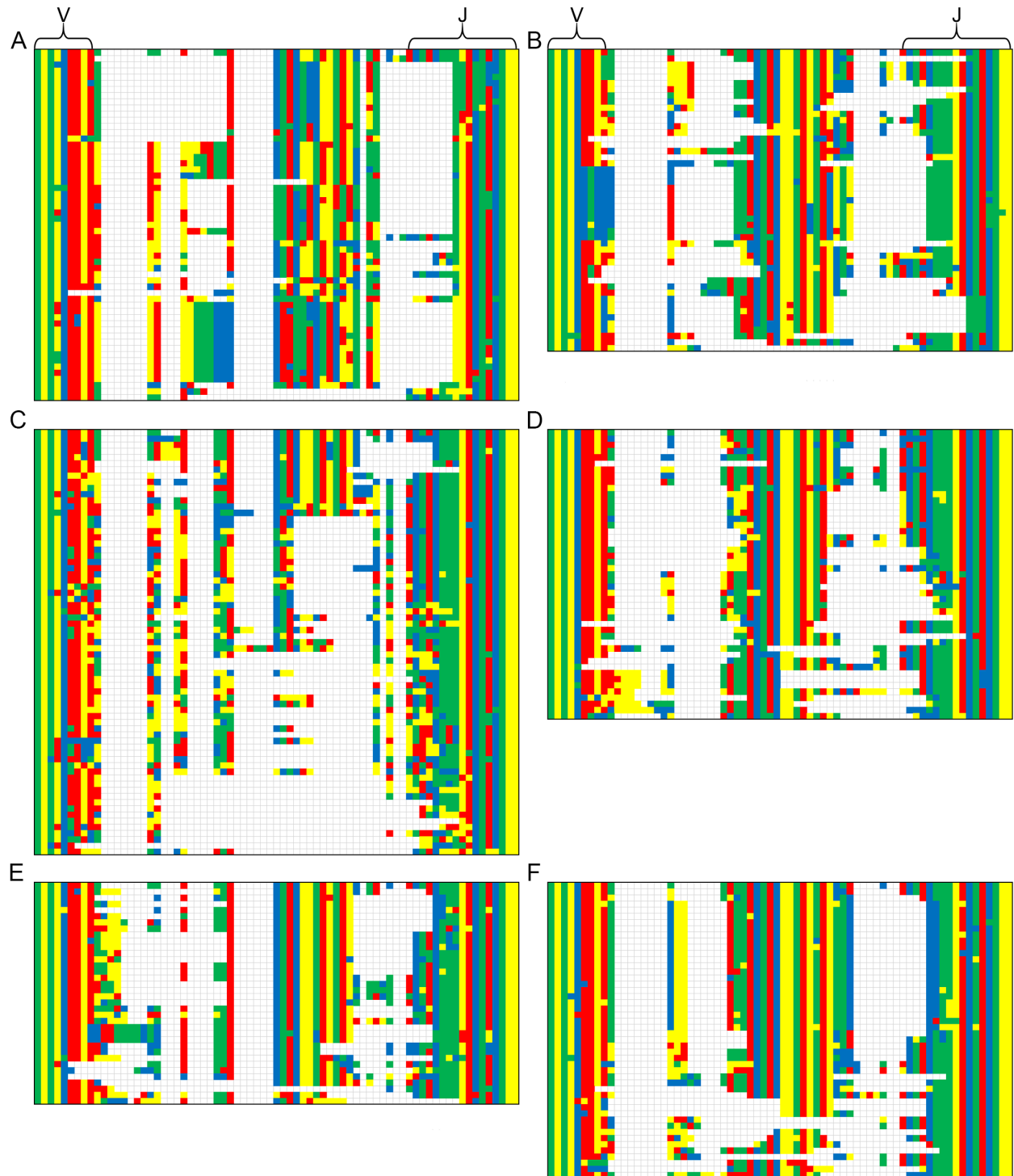
(A) IgH and (B) Igk CDR3 amino acid length of ground control and flight animals as mean-average with standard deviation.

**Figure 4.6 Top CDR3 usage and overlap of CDR3 between treatment animals**



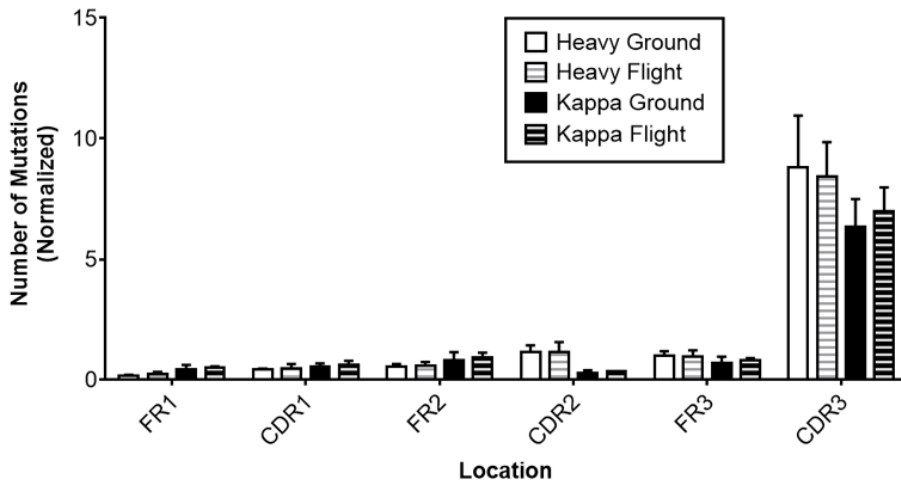
Venn diagrams show the overlap of unique CDR3 sequences among and between ground control (G1-G3) and flight treatment (F1-F3) groups in (A) IgH and (B) Igκ. CDR3 sequences were ranked within the top 5 most abundant sequences of any ground control or flight animals in (C) IgH and (D) Igκ. Dark red indicates higher percent of repertoire moving to blue, which represents lower percent of repertoire.

Figure 4.7 Nucleotide alignment of CDR3 from top V-D-J combination



Nucleotide alignment of heavy-chain gene segment V1-26\*01/D1-1\*01/J1\*03 across individuals in ground (G1 - A, G2 - C, G3 - E) and flight (F1 - B, F2 - D, F3 - F) treatment groups. Brackets in the germline region of the first individual in each treatment group delineate V- and J-gene regions. These bracketed regions remain the same across all individuals in the treatment group.

**Figure 4.8 Substitution mutations by Ig region**



Total number of substitution mutations in FRs 1-3 and CDRs 1-3 were observed for both IgH and Igk chains. Abundance was first normalized by region length and then by total number of cleaned, productive reads in each respective data set and multiplied by 100 to attain percent abundance.

**Table 4.1 Spleen sequencing read counts in ground (G) and flight (F) mice**

	<b>G1</b>	<b>G2</b>	<b>G3</b>	<b>F1</b>	<b>F2</b>	<b>F3</b>
Raw Reads <sup>a,b</sup>	51.4 M	45.2 M	43.5 M	40.5 M	48.4 M	55.8 M
Cleaned <sup>a,c</sup>	13.2 M	31.4 M	30.9 M	14.6 M	13.0 M	14.1 M
IgH IMGT <sup>d,e</sup>	124,102	104,135	149,675	85,802	66,909	181,703
Igκ IMGT <sup>d,e</sup>	172,660	139,777	105,374	108,889	82,653	108,883

<sup>a</sup>M=Million

<sup>b</sup>Raw reads reflect unfiltered FASTQ files imported from the Illumina MiSeq personal sequencing system.

<sup>c</sup>Cleaned reads were quality trimmed to remove the first 12 base pairs, reads with a Phred score under 20, and sequences less than 40 nt in length.

<sup>d</sup>Mapped Ig Sequencing reads of productive or unknown functionality were obtained from the IMGT HighV-Quest tool.

<sup>e</sup>There is no statistical significance by Student's *t*-test in the number of reads mapped to IgH (p=0.7215) and Igκ (p=0.1401).

**Table 4.2 Comparison of flight and ground V-gene segment usage**

Comparison <sup>a</sup>	VH R <sup>2</sup>	Vκ R <sup>2</sup>
G1 v G2	0.616	0.661
G2 v G3	0.338	0.738
G1 v G3	0.686	0.674
F1 v F2	0.110	0.466
F2 v F3	0.227	0.519
F1 v F3	0.367	0.603
Mean G v F	0.583	0.830

Pairwise linear regressions of VH- and Vκ-gene segment usage were performed among ground control (G) and flight (F) animals. A linear regression was performed on the mean-average V gene segment of ground and flight treatment groups.

<sup>a</sup>All comparison groups were correlated ( $p \leq 0.0001$ )



**Table 4.3 V $\kappa$  liver sequencing read counts in ground (G) and flight (F) mice**

	<b>G1</b>	<b>G2</b>	<b>G3</b>	<b>F1</b>	<b>F2</b>	<b>F3</b>
Raw Reads <sup>a,b</sup>	43.9 M	42.9 M	38.4 M	49.2 M	48.3 M	39.7 M
Cleaned <sup>a,c</sup>	18.8 M	35 M	31.5 M	24.7 M	18.2 M	19.4 M
Ig $\kappa$ IMGT <sup>d,e</sup>	1287	1154	471	376	310	309

<sup>a</sup>M=Million

<sup>b</sup>Raw reads reflect unfiltered FASTQ files imported from the Illumina MiSeq personal sequencing system.

<sup>c</sup>Cleaned reads were quality trimmed to remove the first 12 base pairs, reads with a Phred score under 20, and sequences less than 40 nt in length.

<sup>d</sup>Mapped Ig Sequencing reads of productive or unknown functionality were obtained from the IMGT HighV-Quest tool.

**Table 4.4 V-J linear regression analyses**

<b>Comparison<sup>a</sup></b>	<b>IgH R<sup>2</sup></b>	<b>Igκ R<sup>2</sup></b>
G1 vs G2 <sup>a</sup>	0.501	0.751
G2 vs G3 <sup>a</sup>	0.117	0.724
G1 vs G3 <sup>a</sup>	0.202	0.658
F1 vs F2 <sup>b</sup>	0.010	0.601
F2 vs F3 <sup>a</sup>	0.062	0.608
F1 vs F3 <sup>a</sup>	0.103	0.807
G vs F AVG <sup>a</sup>	0.323	0.890

Pairwise linear regression analyses of IgH and Igκ V/J-gene segment combination abundances were performed among grounds control (G) and flight (F) animals. A linear regression was performed on the mean-average V/J-gene segment combination abundances of ground and flight treatment groups.

<sup>a</sup>Comparison group was correlated ( $p < 0.0001$ )

<sup>b</sup>Comparison group was correlated for IgH ( $p = 0.0327$ ) and Igκ ( $p < 0.0001$ )

**Table 4.5 CDR3 length by isotype**

<b>Isotype</b>	<b>CDR3 Amino Acid Length</b>					
	<b>G1</b>	<b>G2</b>	<b>G3</b>	<b>F1</b>	<b>F2</b>	<b>F3</b>
IgA	10	13	12	13	11	12
IgD	11	12	12	12	12	12
IgE	14	11	12	14	11	15
IgG	12	12	13	13	11	12
IgM	12	12	12	12	11	12

The mean-average CDR3 length of ground control (G) and flight (F) animals is displayed by isotype. Assessment by two-way ANOVA revealed no significant differences in CDR3 length by treatment group ( $p=0.6225$ ), isotype ( $p=0.4589$ ), or the interaction of the two variables.

## References

1. Berry, C. A. 1970. Summary of medical experience in the Apollo 7 through 11 manned spaceflights. *Aerospace Med.* 41: 500-519.
2. Crucian, B., R. Stowe, S. Mehta, P. Uchakin, H. Quiariarte, D. Pierson, and C. Sams. 2013. Immune system dysregulation occurs during short duration spaceflight on board the space shuttle. *J. Clin. Immunol.* 33: 456-465.
3. Crucian, B. E., S. R. Zwart, S. Mehta, P. Uchakin, H. D. Quiariarte, D. Pierson, C. F. Sams, and S. M. Smith. 2014. Plasma cytokine concentrations indicate that in vivo hormonal regulation of immunity is altered during long-duration spaceflight. *J. Interferon Cytokine Res.* 34: 778-786.
4. Grigoriev, A. I., S. A. Bugrov, V. V. Bogomolov, A. D. Egorov, V. V. Polyakov, I. K. Tarasov, and E. B. Shulzhenko. 1993. Main medical results of extended flights on space station Mir in 1986-1990. *Acta Astronautica* 29: 581-585.
5. Stein, T. P., and M. D. Schluter. 1994. Excretion of IL-6 by astronauts during spaceflight. *Am. J. Physiol.* 266: E448-452.
6. Taylor, G. R., and J. R. Dardano. 1983. Human cellular immune responsiveness following space flight. *Aviat. Space Environ. Med.* 54: S55-59.
7. Taylor, G. R., L. S. Neale, and J. R. Dardano. 1986. Immunological analyses of U.S. Space Shuttle crewmembers. *Aviat. Space Environ. Med.* 57: 213-217.
8. Rykova, M. P., E. N. Antropova, I. M. Larina, and B. V. Morukov. 2008. Humoral and ceelular immunity in cosmonauts after the ISS missions. *Acta Astronautica* 63: 697-705.

9. Chapes, S. K., A. M. Mastro, G. Sonnenfeld, and W. D. Berry. 1993. Antiorthostatic suspension as a model for the effects of spaceflight on the immune system. *J. Leukoc. Biol.* 54: 227-235.
10. Globus, R. K., and E. Morey-Holton. 2016. Hindlimb unloading: rodent analog for microgravity. *J. Appl. Physiol.* 120: 1196-1206.
11. Crucian, B., R. J. Simpson, S. Mehta, R. Stowe, A. Chouker, S. A. Hwang, J. K. Actor, A. P. Salam, D. Pierson, and C. Sams. 2014. Terrestrial stress analogs for spaceflight associated immune system dysregulation. *Brain Behav. Immun.* 39: 23-32.
12. Nickerson, C. A., C. M. Ott, J. W. Wilson, R. Ramamurthy, C. L. LeBlanc, K. Honer zu Bentrup, T. Hammond, and D. L. Pierson. 2003. Low-shear modeled microgravity: a global environmental regulatory signal affecting bacterial gene expression, physiology, and pathogenesis. *J. Microbiol. Methods* 54: 1-11.
13. Sonnenfeld, G. 2005. Experimentation with animal models in space. Introduction. *Advances in space biology and medicine.* 10: 1-5.
14. Allebban, Z., A. T. Ichiki, L. A. Gibson, J. B. Jones, C. C. Congdon, and R. D. Lange. 1994. Effects of spaceflight on the number of rat peripheral blood leukocytes and lymphocyte subsets. *J. Leukoc. Biol.* 55: 209-213.
15. Chapes, S. K., S. J. Simske, A. D. Forsman, T. A. Bateman, and R. J. Zimmerman. 1999. Effects of space flight and IGF-1 on immune function. *Adv. Space Res.* 23: 1955-1964.
16. Gridley, D. S., J. M. Slater, X. Luo-Owen, A. Rizvi, S. K. Chapes, L. S. Stodieck, V. L. Ferguson, and M. J. Pecaut. 2009. Spaceflight effects on T lymphocyte distribution, function and gene expression. *J. Appl. Physiol.* 106: 194-202.

17. Gridley, D. S., X. W. Mao, L. S. Stodieck, V. L. Ferguson, T. A. Bateman, M. Moldovan, C. E. Cunningham, T. A. Jones, J. M. Slater, and M. J. Pecaut. 2013. Changes in mouse thymus and spleen after return from the STS-135 mission in space. *PloS one* 8: e75097.
18. Ichiki, A. T., L. A. Gibson, T. L. Jago, K. M. Strickland, D. L. Johnson, R. D. Lange, and Z. Allebban. 1996. Effects of spaceflight on rat peripheral blood leukocytes and bone marrow progenitor cells. *J. Leukoc. Biol.* 60: 37-43.
19. Pecaut, M. J., G. A. Nelson, L. L. Peters, P. J. Kostenuik, T. A. Bateman, S. Morony, L. S. Stodieck, D. L. Lacey, S. J. Simske, and D. S. Gridley. 2003. Genetic models in applied physiology: selected contribution: effects of spaceflight on immunity in the C57BL/6 mouse. I. Immune population distributions. *J. Appl. Physiol.* 94: 2085-2094.
20. Sonnenfeld, G., A. D. Mandel, I. V. Konstantinova, G. R. Taylor, W. D. Berry, S. R. Wellhausen, A. T. Lesnyak, and B. B. Fuchs. 1990. Effects of spaceflight on levels and activity of immune cells. *Aviat. Space Environ. Med.* 61: 648-653.
21. Sonnenfeld, G., A. D. Mandel, I. V. Konstantinova, W. D. Berry, G. R. Taylor, A. T. Lesnyak, B. B. Fuchs, and A. L. Rakhmilevich. 1992. Spaceflight alters immune cell function and distribution. *J. Appl. Physiol.* 73: 191s-195s.
22. Gaignier, F., V. Schenten, M. De Carvalho Bittencourt, G. Gauquelin-Koch, J. P. Fripiat, and C. Legrand-Frossi. 2014. Three weeks of murine hindlimb unloading induces shifts from B to T and from th to tc splenic lymphocytes in absence of stress and differentially reduces cell-specific mitogenic responses. *PloS one* 9: e92664.
23. Wei, L. X., J. N. Zhou, A. I. Roberts, and Y. F. Shi. 2003. Lymphocyte reduction induced by hindlimb unloading: distinct mechanisms in the spleen and thymus. *Cell Res.* 13: 465-471.

24. Armstrong, J. W., K. A. Nelson, S. J. Simske, M. W. Luttges, J. J. Iandolo, and S. K. Chapes. 1993. Skeletal unloading causes organ-specific changes in immune cell responses. *J. Appl. Physiol.* 75: 2734-2739.
25. Baqai, F. P., D. S. Gridley, J. M. Slater, X. Luo-Owen, L. S. Stodieck, V. Ferguson, S. K. Chapes, and M. J. Pecaute. 2009. Effects of spaceflight on innate immune function and antioxidant gene expression. *J. Appl. Physiol.* 106: 1935-1942.
26. Chapes, S. K., S. J. Simske, G. Sonnenfeld, E. S. Miller, and R. J. Zimmerman. 1999. Effects of spaceflight and PEG-IL-2 on rat physiological and immunological responses. *J. Appl. Physiol.* 86: 2065-2076.
27. Congdon, C. C., Z. Allebban, L. A. Gibson, A. Kaplansky, K. M. Strickland, T. L. Jago, D. L. Johnson, R. D. Lange, and A. T. Ichiki. 1996. Lymphatic tissue changes in rats flown on Spacelab Life Sciences-2. *J. Appl. Physiol.* 81: 172-177.
28. Durnova, G. N., A. S. Kaplansky, and V. V. Portugalov. 1976. Effect of a 22-day space flight on the lymphoid organs of rats. *Avt. Space Environ. Med.* 47: 588-591.
29. Grove, D. S., S. A. Pishak, and A. M. Mastro. 1995. The effect of a 10-day space flight on the function, phenotype, and adhesion molecule expression of splenocytes and lymph node lymphocytes. *Exp. Cell Res.* 219: 102-109.
30. Pecaute, M. J., S. J. Simske, and M. Fleshner. 2000. Spaceflight induces changes in splenocyte subpopulations: effectiveness of ground-based models. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 279: R2072-2078.
31. Ortega, M. T., M. J. Pecaute, D. S. Gridley, L. S. Stodieck, V. Ferguson, and S. K. Chapes. 2009. Shifts in bone marrow cell phenotypes caused by spaceflight. *J. Appl. Physiol.* 106: 548-555.

32. Lescale, C., V. Schenten, D. Djeghloul, M. Bennabi, F. Gaignier, K. Vandamme, C. Strazielle, I. Kuzniak, H. Petite, C. Dosquet, J. P. Fripiat, and M. Goodhardt. 2015. Hind limb unloading, a model of spaceflight conditions, leads to decreased B lymphopoiesis similar to aging. *FASEB J.* 29: 455-463.
33. Cooper, D., M. W. Pride, E. L. Brown, D. Risin, and N. R. Pellis. 2001. Suppression of antigen-specific lymphocyte activation in modeled microgravity. *In Vitro Cell. Dev. Biol. Anim.* 37: 63-65.
34. Lesnyak, A. T., G. Sonnenfeld, M. P. Rykova, D. O. Meshkov, A. Mastro, and I. Konstantinova. 1993. Immune changes in test animals during spaceflight. *J. Leukoc. Biol.* 54: 214-226.
35. Lesnyak, A., G. Sonnenfeld, L. Avery, I. Konstantinova, M. Rykova, D. Meshkov, and T. Orlova. 1996. Effect of SLS-2 spaceflight on immunologic parameters of rats. *J. Appl. Physiol.* 81: 178-182.
36. Nash, P. V., I. V. Konstantinova, B. B. Fuchs, A. L. Rakhmievich, A. T. Lesnyak, and A. M. Mastro. 1992. Effect of spaceflight on lymphocyte proliferation and interleukin-2 production. *J. Appl. Physiol.* 73: 186s-190s.
37. Nash, P. V., and A. M. Mastro. 1992. Variable lymphocyte responses in rats after space flight. *Exp. Cell. Res.* 202: 125-131.
38. Sanzari, J. K., A. L. Romero-Weaver, G. James, G. Krigsfeld, L. Lin, E. S. Diffenderfer, and A. R. Kennedy. 2013. Leukocyte activity is altered in a ground based murine model of microgravity and proton radiation exposure. *PloS one* 8: e71757.



39. Sonnenfeld, G., M. Foster, D. Morton, F. Bailliard, N. A. Fowler, A. M. Hakenewerth, R. Bates, and E. S. Miller, Jr. 1998. Spaceflight and development of immune responses. *J. Appl. Physiol.* 85: 1429-1433.
40. Cogoli, A., A. Tschopp, and P. Fuchs-Bislin. 1984. Cell sensitivity to gravity. *Science* 225: 228-230.
41. Cogoli-Greuter, M. 2004. Effect of Gravity Changes on the Cytoskeleton in Human Lymphocytes. *Gravitat. Space Biol. Bull.* 17: 27-38.
42. Chang, T. T., I. Walther, C. F. Li, J. Boonyaratanakornkit, G. Galleri, M. A. Meloni, P. Pippia, A. Cogoli, and M. Hughes-Fulford. 2012. The Rel/NF-kappaB pathway and transcription of immediate early genes in T cell activation are inhibited by microgravity. *J. Leukoc. Biol.* 92: 1133-1145.
43. Hwang, S. A., B. Crucian, C. Sams, and J. K. Actor. 2015. Post-Spaceflight (STS-135) Mouse Splenocytes Demonstrate Altered Activation Properties and Surface Molecule Expression. *PloS one* 10: e0124380.
44. Martinez, E. M., M. C. Yoshida, T. L. Candelario, and M. Hughes-Fulford. 2015. Spaceflight and simulated microgravity cause a significant reduction of key gene expression in early T-cell activation. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 308: R480-488.
45. Tauber, S., S. Hauschild, K. Paulsen, A. Gutewort, C. Raig, E. Hurlimann, J. Biskup, C. Philpot, H. Lier, F. Engelmann, A. Pantaleo, A. Cogoli, P. Pippia, L. E. Layer, C. S. Thiel, and O. Ullrich. 2015. Signal transduction in primary human T lymphocytes in altered gravity during parabolic flight and clinostat experiments. *Cell. Physiol. Biochem.* 35: 1034-1051.

46. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575-581.
47. Early, P., H. Huang, M. Davis, K. Calame, and L. Hood. 1980. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. *Cell* 19: 981-992.
48. Sakano, H., K. Huppi, G. Heinrich, and S. Tonegawa. 1979. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* 280: 288-294.
49. Hozumi, N., and S. Tonegawa. 1976. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl. Acad. Sci.* 73: 3628-3632.
50. Alt, F. W., G. D. Yancopoulos, T. K. Blackwell, C. Wood, E. Thomas, M. Boss, R. Coffman, N. Rosenberg, S. Tonegawa, and D. Baltimore. 1984. Ordered rearrangement of immunoglobulin heavy chain variable region segments. *The EMBO journal* 3: 1209-1219.
51. Gilfillan, S., A. Dierich, M. Lemeur, C. Benoist, and D. Mathis. 1993. Mice lacking TdT: mature animals with an immature lymphocyte repertoire. *Science* 261: 1175-1178.
52. Komori, T., A. Okada, V. Stewart, and F. W. Alt. 1993. Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes. *Science* 261: 1171-1175.
53. Kabat, E. A., T. T. Wu, and H. Bilofsky. 1979. Evidence supporting somatic assembly of the DNA segments (minigenes), coding for the framework, and complementarity-determining segments of immunoglobulin variable regions. *J. Exp. Med.* 149: 1299-1313.
54. Xu, J. L., and M. M. Davis. 2000. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* 13: 37-45.

55. Georgiou, G., G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnol.* 32: 158-168.
56. Rettig, T. A., C. Ward, M. J. Pecaut, and S. K. Chapes. 2017. Validation of Methods to Assess the Immunoglobulin Gene Repertoire in Tissues Obtained from Mice on the International Space Station. *Gravitat. Space Res.* 5: 2-23.
57. Wardemann, H. B., C.E. 2017. Novel Approaches to Analyze Immunoglobulin Repertoires. *Trends Immunol.* 38: 471-482.
58. Ademokun, A., Y. C. Wu, V. Martin, R. Mitra, U. Sack, H. Baxendale, D. Kipling, and D. K. Dunn-Walters. 2011. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* 10: 922-930.
59. Khurana, S., S. Fuentes, E. M. Coyle, S. Ravichandran, R. T. Davey, Jr., and J. H. Beigel. 2016. Human antibody repertoire after VSV-Ebola vaccination identifies novel targets and virus-neutralizing IgM antibodies. *Nature Med.* 22: 1439-1447.
60. Lee, J., D. R. Boutz, V. Chromikova, M. G. Joyce, C. Vollmers, K. Leung, A. P. Horton, B. J. DeKosky, C. H. Lee, J. J. Lavinder, E. M. Murrin, C. Chrysostomou, K. H. Hoi, Y. Tsybovsky, P. V. Thomas, A. Druz, B. Zhang, Y. Zhang, L. Wang, W. P. Kong, D. Park, L. I. Popova, C. L. Dekker, M. M. Davis, C. E. Carter, T. M. Ross, A. D. Ellington, P. C. Wilson, E. M. Marcotte, J. R. Mascola, G. C. Ippolito, F. Krammer, S. R. Quake, P. D. Kwong, and G. Georgiou. 2016. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nature Med.* 22: 1456-1464.

61. Parameswaran, P., Y. Liu, K. M. Roskin, K. K. Jackson, V. P. Dixit, J. Y. Lee, K. L. Artiles, S. Zompi, M. J. Vargas, B. B. Simen, B. Hanczaruk, K. R. McGowan, M. A. Tariq, N. Pourmand, D. Koller, A. Balmaseda, S. D. Boyd, E. Harris, and A. Z. Fire. 2013. Convergent antibody signatures in human dengue. *Cell Host Microbe* 13: 691-700.
62. Tan, Y. C., S. Kongpachith, L. K. Blum, C. H. Ju, L. J. Lahey, D. R. Lu, X. Cai, C. A. Wagner, T. M. Lindstrom, J. Sokolove, and W. H. Robinson. 2014. Barcode-enabled sequencing of plasmablast antibody repertoires in rheumatoid arthritis. *Arthritis Rheumatol. (Hoboken, N.J.)* 66: 2706-2715.
63. Tan, Y. G., Y. Q. Wang, M. Zhang, Y. X. Han, C. Y. Huang, H. P. Zhang, Z. M. Li, X. L. Wu, X. F. Wang, Y. Dong, H. M. Zhu, S. D. Zhu, H. M. Li, N. Li, H. P. Yan, and Z. H. Gao. 2016. Clonal Characteristics of Circulating B Lymphocyte Repertoire in Primary Biliary Cholangitis. *J. Immunol.* 197: 1609-1620.
64. Zuckerman, N. S., W. A. Howard, J. Bismuth, K. Gibson, H. Edelman, S. Berrih-Aknin, D. Dunn-Walters, and R. Mehr. 2010. Ectopic GC in the thymus of myasthenia gravis patients show characteristics of normal GC. *Eur. J. Immunol.* 40: 1150-1161.
65. Bashford-Rogers, R. J., K. A. Nicolaou, J. Bartram, N. J. Goulden, L. Loizou, L. Koumas, J. Chi, M. Hubank, P. Kellam, P. A. Costeas, and G. S. Vassiliou. 2016. Eye on the B-ALL: B-cell receptor repertoires reveal persistence of numerous B-lymphoblastic leukemia subclones from diagnosis to relapse. *Leukemia* 30: 2312-2321.
66. Jiang, Y., K. Nie, D. Redmond, A. M. Melnick, W. Tam, and O. Elemento. 2015. VDJ-Seq: Deep Sequencing Analysis of Rearranged Immunoglobulin Heavy Chain Gene to Reveal Clonal Evolution Patterns of B Cell Lymphoma. *JoVE*: e53215.

67. Logan, A. C., H. Gao, C. Wang, B. Sahaf, C. D. Jones, E. L. Marshall, I. Buno, R. Armstrong, A. Z. Fire, K. I. Weinberg, M. Mindrinos, J. L. Zehnder, S. D. Boyd, W. Xiao, R. W. Davis, and D. B. Miklos. 2011. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc. Natl. Acad. Sci.* 108: 21194-21199.
68. Montesinos-Rongen, M., F. Purschke, R. Kuppers, and M. Deckert. 2014. Immunoglobulin repertoire of primary lymphomas of the central nervous system. *J. Neuropathol. Exp. Neurol.* 73: 1116-1125.
69. Tschumper, R. C., Y. W. Asmann, A. Hossain, P. M. Huddleston, X. Wu, A. Dispenzieri, B. W. Eckloff, and D. F. Jelinek. 2012. Comprehensive assessment of potential multiple myeloma immunoglobulin heavy chain V-D-J intracлонаl variation using massively parallel pyrosequencing. *Oncotarget* 3: 502-513.
70. Greiff, V., P. Bhat, S. C. Cook, U. Menzel, W. Kang, and S. T. Reddy. 2015. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* 7: 49.
71. Fitzgerald, W., S. Chen, C. Walz, J. Zimmerberg, L. Margolis, and J. C. Grivel. 2009. Immune suppression of human lymphoid tissues and cells in rotating suspension culture and onboard the International Space Station. *In Vitro Cell. Dev. Biol. Anim.* 45: 622-632.
72. Stowe, R. P., C. F. Sams, S. K. Mehta, I. Kaur, M. L. Jones, D. L. Feedback, and D. L. Pierson. 1999. Leukocyte subsets and neutrophil function after short-term spaceflight. *J. Leukoc. Biol.* 65: 179-186.
73. Voss, E. W., Jr. 1984. Prolonged weightlessness and humoral immunity. *Science* 225: 214-215.

74. Boxio, R., C. Dournon, and J. P. Frippiat. 2005. Effects of a long-term spaceflight on immunoglobulin heavy chains of the urodele amphibian *Pleurodeles waltl*. *J. Appl. Physiol.* 98: 905-910.
75. Bascove, M., C. Huin-Schohn, N. Gueguinou, E. Tschirhart, and J. P. Frippiat. 2009. Spaceflight-associated changes in immunoglobulin VH gene expression in the amphibian *Pleurodeles waltl*. *FASEB J.* 23: 1607-1615.
76. Bascove, M., N. Gueguinou, B. Schaerlinger, G. Gauquelin-Koch, and J. P. Frippiat. 2011. Decrease in antibody somatic hypermutation frequency under extreme, extended spaceflight conditions. *FASEB J.* 25: 2947-2955.
77. Huin-Schohn, C., N. Gueguinou, V. Schenten, M. Bascove, G. G. Koch, S. Baatout, E. Tschirhart, and J. P. Frippiat. 2013. Gravity changes during animal development affect IgM heavy-chain transcription and probably lymphopoiesis. *FASEB J.* 27: 333-341.
78. Alamyar, E., P. Duroux, M. P. Lefranc, and V. Giudicelli. 2012. IMGT((R)) 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.* 882: 569-604.
79. Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res.* 19: 1639-1645.
80. Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30: 3059-3066.

81. Loder, F., B. Mutschler, R. J. Ray, C. J. Paige, P. Sideras, R. Torres, M. C. Lamers, and R. Carsetti. 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: 75-89.
82. Yang, Y., C. Wang, Q. Yang, A. B. Kantor, H. Chu, E. E. Ghosn, G. Qin, S. K. Mazmanian, J. Han, and L. A. Herzenberg. 2015. Distinct mechanisms define murine B cell lineage immunoglobulin heavy chain (IgH) repertoires. *eLife* 4: e09083.
83. Kaplinsky, J., A. Li, A. Sun, M. Coffre, S. B. Koralov, and R. Arnaout. 2014. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc. Natl. Acad. Sci.* 111: E2622-2629.
84. Menzel, U., V. Greiff, T. A. Khan, U. Haessler, I. Hellmann, S. Friedensohn, S. C. Cook, M. Pogson, and S. T. Reddy. 2014. Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PloS one* 9: e96727.
85. Ghislin, S., N. Ouzren-Zarhloul, S. Kaminski, and J. P. Frippiat. 2015. Hypergravity exposure during gestation modifies the TCRbeta repertoire of newborn mice. *Sci. Rep.* 5: 9318.
86. Schenten, V., N. Gueguinou, S. Baatout, and J. P. Frippiat. 2013. Modulation of *Pleurodeles waltl* DNA polymerase mu expression by extreme conditions encountered during spaceflight. *PloS one* 8: e69647.
87. Garcia, K. C., C. A. Scott, A. Brunmark, F. R. Carbone, P. A. Peterson, I. A. Wilson, and L. Teyton. 1996. CD8 enhances formation of stable T-cell receptor/MHC class I molecule complexes. *Nature* 384: 577-581.

88. Cancro, M. P. H., Y., Scholz, J.L.; Riley, R.L.; Frasca, D.; Dunn-Walters, D.K.; Blomberg, B.B. 2009. B cells and aging: molecules and mechanisms. *Trends Immunol.* 30: 313-318.



## Chapter 5 - Conclusions

This work characterizes the Ig $\kappa$  repertoire from high throughput sequencing datasets to test whether space flight affects the immunoglobulin repertoire of unimmunized C57BL/6 mice.

In the second chapter, sample preparation methods and bioinformatics workflows were validated. While traditional immunoglobulin repertoire studies assess sorted B-cell populations, the isolation of sorted B-cell populations is not practical in a spaceflight environment. Immunoglobulin repertoires sequenced from pooled splenic single cell suspensions and whole tissue were compared. Immunoglobulin repertoires from single cell suspensions, whole tissue and a size selected subset of whole tissue from a single mouse pool were uniform in gene segment usage. Highly abundant Ig $\kappa$ -gene segments were conserved in all three treatment grounds. Despite similarity in gene segment usage,  $\kappa$ -CDR3 sequence composition was varied among cells, tissue and size selected tissue treatment groups, demonstrating the combinatorial diversity of the immunoglobulin repertoire.

Second, traditional immunoglobulin repertoires rely on the specific amplification of immunoglobulin sequences to ensure the capture of rare immunoglobulins. Because this practice precludes the possibility of mining datasets for additional information, no amplification of immunoglobulin sequences was used in the generation of the HTS datasets. Without amplification, 10,000-24,000 Ig $\kappa$  sequencing reads were recovered. The size-selected whole tissue treatment group provided the highest yield of Ig $\kappa$  sequences, therefore this sample preparation method was used for subsequent experiments.

Finally, characterization of the immunoglobulin repertoire is enriched when long sequencing reads are generated as more information on gene segment usage, junction and constant region identity can be obtained. Modifications in the bioinformatic workflow were employed to

ensure fidelity in publically available datasets that contained shorter sequencing reads. Without modification, the amplification of our standard bioinformatic workflow of mapping to the Ig $\kappa$  locus and V $\kappa$ -gene segments to shorter HTS reads from unimmunized mouse datasets resulted in poor correlation to the immunoglobulin repertoire of unimmunized mouse HTS datasets with longer sequencing reads. After mapping shorter reads to the entire mouse genome and selecting reads that mapped to the Ig $\kappa$  locus, immunoglobulin repertoire correlation between short and long HTS reads from unimmunized mice was restored.

In the third chapter, immunoglobulin repertoires of pooled biological replicates were characterized and compared. Patterns of V-gene segment usage were assessed, revealing non-random V-gene segment usage that showed bias by chromosomal location. Gene segment usage was conserved among pooled samples, however gene segment combinations and CDR3 were more diverse. Overlap of gene segment combinations and CDR3 use was found when comparing mouse pools. Still, a large number of these sequences were unique within each mouse pool, demonstrating the diversity within the immunoglobulin repertoire. CDR3 amino acid sequence length in Ig $\kappa$  was conserved at nine amino acids. Junctional diversity within Ig $\kappa$  was illustrated through the nucleotide alignment single V- and J-gene segment pair, resulting in 30 unique amino acid sequences. Overall, CDR3 within Ig $\kappa$  were more homogeneous than H-CDR3. Importantly, the data for pool one was used in the second chapter to validate sample preparation and the bioinformatics workflow. Pools two and three were prepared from size selected tissue, as the sample preparation assessment revealed that this method yielded the highest number of immunoglobulin sequencing reads. Pools two and three yielded even more immunoglobulin sequences, validating the sample preparation method.

In the fourth chapter, the immunoglobulin repertoires of unimmunized mice flown aboard the International Space Station were compared to the immunoglobulin repertoires of unimmunized ground control mice. The hypothesis that spaceflight impacts the Igk repertoire was tested by comparing V- and J-gene segment usage, and CDR3 diversity. Variation between individual animals was high and no significant differences were found in gene segment usage between flight and ground treatment groups. Additionally, Illumina MiSeq sequencing was performed on both liver and spleen samples, however, too few immunoglobulin transcripts were found within the liver for the assessment of tissue specific immunoglobulin repertoires.

Overall, individual gene segment usage was well correlated between the sample preparation treatment groups prepared from a single mouse pool (Pool 1) in chapter 2 (Cells, Tissue, and Size Selected), pooled replicates in chapter 3 (Pools 1, 2, and 3), and individual RR1 mice in chapter 4 (Ground and Flight). The combinatorial diversity of the immunoglobulin repertoire was seen through the variation CDR3 usage in both three pooled mouse datasets from the chapters 2 and 3 and individual mouse datasets used in chapter 4, with a large number of CDR3 sequences being identified in only one dataset within each study. In toto, this work has shown the successful development of methodology to assess Ig repertoires in mice. The methodology was validated in two different studies and we were able to successfully assess normal mouse Ig repertoires in several different data sets. The data collected allowed us to compare two different normal mouse groups and provoked questions about how age and origin of mice affects the normal Ig repertoire. In addition, this work has prepared us for future studies to assess antibody repertoire dynamics in response to vaccination within the setting of a spaceflight analog and space flight.

## Appendix A - Specific Procedures for Repertoire Analysis

### Appendix A.1 Statement of copyright release

#### Copyright Release Request

Copyright release for the material specified below has been requested for the inclusion as a chapter in the Masters of Science thesis of the Kansas State University student Claire Ward.

Manuscript Title: Validation of Methods to Assess the Immunoglobulin Gene Repertoire in Tissues Obtained from Mice on the International Space Station

Authors: Trisha A. Rettig, Claire Ward, Michael J. Pecaut, Stephen K. Chapes

Journal: Gravitational and Space Research

Publication Date: In Press, Submitted November 2016

#### Statement of Copyright Release

I hereby represent that I have the authority to grant the permission requested herein.

I grant copyright release for the requested material to be used for the above specified purpose.

<u>Dr. Anna-Lisa Paul</u>	
Name of authorized signatory (Print)	
<u>Research Professor / Editor in Chief GSR</u>	<u>University of Florida / ASGSR</u>
Title	Institution
<u>2550 Hull Rd, Gainesville, FL, 32611</u>	<u>alp@ufl.edu</u>
Address	Email
	<u>23 July 2017</u>
Signature of authorized signatory	Date

## Appendix A.2 V-gene segment usage in normal mouse pools

A

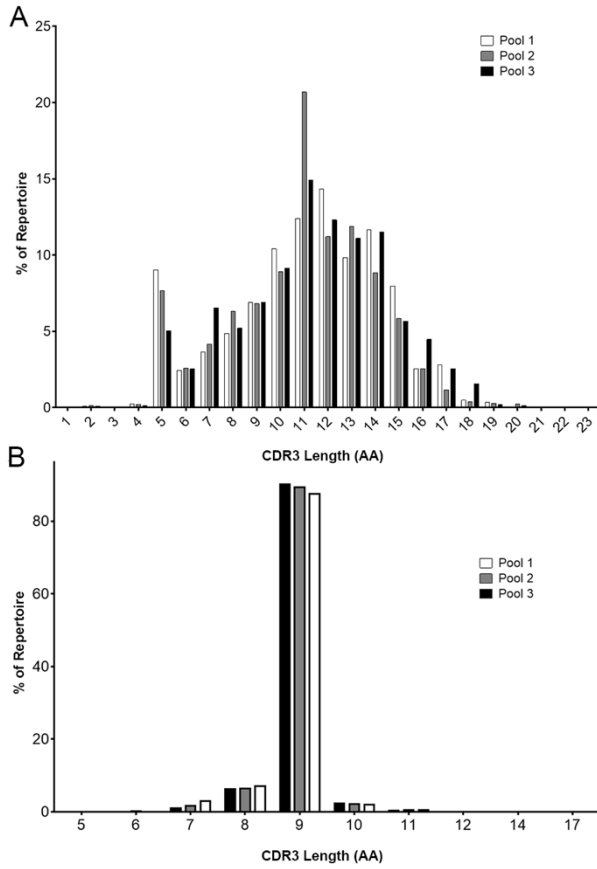
	Pool 1	Pool 2	Pool 3		Pool 1	Pool 2	Pool 3		Pool 1	Pool 2	Pool 3
V1-80*01	7.75	6.87	3.31	V5-6*01	0.39	0.37	0.56	V8-6-1*01	0.01	0.02	0.02
V1-26*01	4.29	3.95	3.97	V1-5*01	0.25	0.51	0.42	V5-12-4*01	0.01	0.01	0.02
V6-3*01	2.91	2.74	5.06	V1-84*01	0.50	0.26	0.42	V8-9*01	0.01	0.01	0.02
V1-53*01	3.79	2.95	2.52	V8-5*01	0.43	0.32	0.40	V5-1*01	0.00	0.02	0.02
V1-55*01	3.57	2.40	2.92	V13-2*01	0.59	0.24	0.30	V5-21*01	0.01	0.01	0.02
V3-6*01	2.68	2.27	3.91	V14-1*01	0.32	0.52	0.28	V1-27*01	0.01	0.01	0.02
V1-9*01	2.99	3.54	1.90	V9-2*01	0.14	0.32	0.58	V1-21-1*01	0.01	0.02	0.01
V1-18*01	4.04	3.08	1.24	V5S21*01	0.07	0.11	0.72	V1-62-1*01	0	0.01	0.02
V1-50*01	1.43	2.32	3.99	V1-47*01	0.33	0.28	0.26	V1-60*01	0.01	0.01	0.01
V9-3*01	2.45	2.72	2.53	V1-34*01	0.30	0.22	0.32	V6-5*01	0.01	0.01	0.01
V1-64*01	1.88	2.08	3.74	V3-5*01	0.27	0.33	0.22	V3-7*01	0	0.01	0.01
V6-6*01	2.72	2.22	1.55	V2-4*01	0.27	0.39	0.15	V1-32*01	0.00	0.01	0.01
V1-82*01	2.06	1.96	2.01	V1-58*01	0.34	0.29	0.17	V1-17-1*01	0.01	0.01	0.00
V8-8*01	2.22	1.74	1.99	V1-20*01	0.19	0.31	0.24	V12-2*01	0	0.02	0.00
V1-78*01	2.66	1.17	1.55	V12-3*01	0.30	0.25	0.17	V5S24*01	0.01	0.01	0.00
V7-3*01	1.67	1.63	1.49	V5-12*01	0.17	0.16	0.31	V1-48*01	0.00	0.00	0.01
V1-81*01	1.75	1.76	1.27	V1-77*01	0.20	0.19	0.20	V1-13*01	0.01	0.00	0.01
V1-72*01	1.31	1.40	1.96	V1S96*01	0.01	0.18	0.37	V1-24*01	0	0.00	0.00
V10-1*01	1.62	1.62	1.41	V1-71*01	0.13	0.10	0.33	V1-19-1*01	0	0.00	0.00
V5-17*01	1.39	1.76	1.41	V5-9*01	0.23	0.18	0.14	V1-35*01	0	0	0.00
V1-22*01	1.50	1.76	1.28	V1-36*01	0.16	0.23	0.15				
V2-2*01	1.16	1.75	1.43	V1-62-2*01	0.13	0.10	0.20				
V14-4*01	1.50	1.29	1.47	V1-11*01	0.11	0.23	0.09				
V1-76*01	0.87	2.22	1.14	V5-15*01	0.08	0.24	0.08				
V2-3*01	1.54	1.40	1.28	V5-2*01	0.15	0.12	0.11				
V4-1*01	1.09	2.05	1.05	V8-2*01	0.09	0.14	0.12				
V1S92*01	1.43	0.99	1.53	V1S5*01	0	0.04	0.30				
V11-2*01	1.59	1.38	0.89	V1-67*01	0.07	0.14	0.11				
V1S95*01	1.34	0.97	1.54	V7-4*01	0.18	0.07	0.06				
V1-69*01	0.86	1.40	1.44	V1-8*01	0.07	0.12	0.12				
V1-39*01	1.03	0.85	1.40	V3-4*01	0.14	0.08	0.09				
V1-15*01	1.04	0.96	1.08	V1-62-3*01	0.02	0.02	0.25				
V10-3*01	0.78	1.47	0.82	V1S87*01	0.00	0.11	0.18				
V2-9*01	1.39	0.82	0.85	V8S9*01	0.03	0.10	0.15				
V1-75*01	1.30	0.72	0.91	V1S103*01	0.00	0.06	0.21				
V2-9-1*01	0.93	0.78	1.16	V1S65*01	0.06	0.07	0.13				
V1-19*01	0.74	1.07	1.04	V1S-2*01	0.07	0.07	0.11				
V14-2*01	1.03	0.99	0.82	V1-49*01	0.05	0.09	0.11				
V8-12*01	1.03	1.00	0.80	V1-31*01	0.08	0.07	0.08				
V1-66*01	0.35	0.47	1.97	V1S107*01	0.04	0.07	0.11				
V1-52*01	0.77	0.98	1.00	V11-1*01	0.01	0.12	0.06				
V5-4*01	0.99	0.91	0.67	V3-3*01	0.09	0.07	0.03				
V1S108*01	0.13	0.74	1.62	V1S100*01	0.02	0.04	0.11				
V5-16*01	0.89	0.83	0.75	V1S110*01	0.03	0.04	0.09				
V1-7*01	0.80	0.88	0.78	V1-23*01	0.05	0.04	0.06				
V5-9-1*02	0.31	0.30	1.78	V1-43*01	0.03	0.05	0.06				
V1-74*01	0.45	0.62	1.21	V8-11*01	0.01	0.05	0.08				
V2-6*01	0.02	1.97	0.24	V2-7*01	0.04	0.06	0.04				
V2-6-8*01	0.02	1.97	0.24	V1-37*01	0.06	0.04	0.03				
V9-1*01	0.43	0.79	0.95	V16-1*01	0.03	0.05	0.04				
V1-59*01	0.90	0.53	0.68	V1S101*01	0.01	0.04	0.06				
V1-63*01	1.68	0.23	0.16	V6-7*01	0.04	0.02	0.05				
V1-42*01	0.77	0.44	0.83	V1S67*01	0.01	0.02	0.07				
V9-4*01	1.06	0.52	0.42	V1S68*01	0.01	0.03	0.06				
V3-1*01	0.92	0.44	0.53	V8S6*01	0.02	0.03	0.06				
V14-3*01	0.60	0.58	0.70	V7-2*01	0.03	0.04	0.03				
V1-54*01	0.82	0.46	0.47	V1-56*01	0.03	0.04	0.03				
V1-4*01	0.56	0.67	0.44	V3S7*01	0.01	0.03	0.04				
V2-5*01	0.39	0.67	0.56	V1-51*01	0.03	0.03	0.03				
V3-8*01	0.45	0.66	0.42	V8-6*01	0.00	0.02	0.05				
V1-12*01	0.54	0.62	0.35	V8-4*01	0.02	0.03	0.02				
V1-61*01	0.44	0.60	0.40	V6-4*01	0.01	0.02	0.03				
V1-85*01	0.67	0.30	0.47	V1-14*01	0.01	0.02	0.02				

B

	Pool 1	Pool 2	Pool 3		Pool 1	Pool 2	Pool 3
V1-117*01	7.96	4.79	4.48	V1-88*01	0.26	0.41	0.28
V1-110*01	4.70	8.22	3.81	V4-50*01	0.31	0.28	0.34
V5-39*01	10.05	1.25	3.46	V6-14*01	0.38	0.23	0.28
V4-55*01	4.14	4.62	4.16	V4-74*01	0.33	0.30	0.22
V1-135*01	3.52	3.15	3.52	V4-80*01	0.21	0.30	0.31
V3-4*01	3.33	2.55	3.32	V7-33*01	0.18	0.30	0.34
V10-96*01	2.60	3.20	3.05	V4-58*01	0.14	0.18	0.22
V6-15*01	1.47	2.24	3.61	V11-125*01	0.18	0.13	0.15
V12-44*01	1.64	3.81	1.73	V1-133*01	0.14	0.13	0.16
V5-43*01	3.24	2.00	1.89	V9-129*01	0.11	0.17	0.11
V3-2*01	2.31	1.46	3.25	V4-71*01	0.03	0.08	0.27
V14-111*01	1.93	2.71	2.37	V3-1*01	0.00	0.18	0.19
V2-137*01	1.94	1.79	3.18	V4-69*01	0.10	0.09	0.17
V8-30*01	1.55	2.31	2.96	V10-95*01	0.15	0.11	0.09
V19-93*01	2.10	2.77	1.93	V18-36*01	0.32	0.01	0.01
V12-46*01	2.17	2.80	1.62	V8-18*01	0.08	0.04	0.22
V9-120*01	1.71	1.94	2.22	V4-90*01	0.15	0.08	0.08
V8-24*01	2.18	1.95	1.70	V9-123*01	0.09	0.10	0.10
V14-59*01	1.89	1.56	2.11	V12-38*01	0.11	0.09	0.08
V4-57-1*01	2.52	1.26	1.41	V8-34*01	0.05	0.12	0.09
V6-23*01	1.02	2.02	1.90	V14-130*01	0.08	0.09	0.06
V6-17*01	1.41	1.85	1.62	V4-81*01	0.08	0.08	0.04
V14-126-1*01	1.58	1.66	1.50	V1-131*01	0.02	0.03	0.15
V15-103*01	0.98	2.24	1.49	V8-23-1*01	0.06	0.04	0.09
V4-70*01	0.65	1.20	2.69	V4-51*01	0.05	0.07	0.05
V10-94*01	0.94	2.22	1.30	V5-37*01	0.11	0.03	0.03
V4-72*01	2.02	1.11	1.30	V4-92*01	0.07	0.05	0.05
V8-27*01	1.35	1.56	1.27	V4-78*01	0.05	0.05	0.05
V9-124*01	1.88	1.15	0.75	V4-54*01	0.06	0.04	0.04
V16-104*01	0.98	1.11	1.35	V4-62*01	0.05	0.01	0.05
V6-32*01	0.81	1.31	1.14	V4-73*01	0.03	0.03	0.03
V2-109*01	0.96	1.16	1.10	V1-132*01	0.03	0.03	0.02
V17-127*01	1.04	1.01	1.14	V3-3*01	0.04	0.02	0.03
V8-21*01	0.36	0.74	2.04	V3-9*01	0.03	0.02	0.03
V4-53*01	1.00	1.31	0.82	V8-26*01	0.01	0.01	0.01
V8-19*01	0.84	1.08	0.94	V20-101-2*01	0.01	0.01	0.01
V6-25*01	0.75	0.62	1.46	V1-35*01	0.02	0.00	0.00
V17-121*01	0.95	0.80	0.94				
V12-89*01	0.93	0.78	0.93				
V4-86*01	0.49	1.22	0.93				
V3-12-1*01	0.51	0.99	0.93				
V12-41*01	0.65	0.59	1.11				
V4-68*01	0.57	1.01	0.74				
V4-63*01	0.44	1.60	0.27				
V4-91*01	0.67	0.68	0.90				
V3-5*01	0.75	0.60	0.91				
V2-112*01	0.38	1.24	0.48				
V3-10*01	0.53	0.65	0.92				
V8-28*01	0.78	0.60	0.62				
V6-20*01	0.53	0.90	0.54				
V14-100*01	0.63	0.64	0.57				
V6-13*01	0.30	0.35	1.02				
V13-85*01	0.49	0.37	0.61				
V3-7*01	0.26	0.67	0.49				
V13-84*01	0.51	0.43	0.43				
V6-29*01	1.15	0.03	0.09				
V5-45*01	0.64	0.20	0.37				
V8-16*01	0.27	0.52	0.36				
V1-122*01	0.47	0.31	0.29				
V12-96*01	0.21	0.54	0.32				
V4-79*01	0.19	0.19	0.67				
V1-99*01	0.56	0.21	0.27				
V4-61*01	0.56	0.23	0.23				

Complete V-gene segment rankings among the three mouse pools for both IgH (A) and Igk (B). The most abundant gene segment is ranked as 1. Dark red indicates higher rank moving to blue, of lower rank.

### Appendix A.3 CDR3 length by individual mouse pool



CDR3 length for IgH (A) and Igκ (B) by pool. The percent of repertoire for CDR3 lengths from each mouse pool is displayed.

## Appendix A.4 CDR3 sequences shared among all pools

Rankings of CDR3 sequences shared by all three mouse pools were uniform in both IgH (A) and Igk (B). The most abundant CDR3 sequence is ranked as 1. Dark red indicates higher rank moving to blue, of lower rank.

**A**

	Pool 1	Pool 2	Pool 3
CARGAYW	1	2	3
CARDYYGSSWYFDVW	18	6	6
CMRYSNYWYFDVW	4	11	15
CARGGYW	16	5	18
CARGTYW	45	18	7
CMRYGNYWYFDVW	37	15	23
CARGYFDYW	45	9	22
CARGDYW	9	26	45
CARSDNWYFDVW	39	34	19
CARGPYW	13	158	34
CARDNWDWYFDVW	189	20	55
CMRYSSWYFDVW	189	14	62
CARW	125	73	170
CARGGFAYW	97	158	137
CAKKGAMDYW	97	79	299
CARDYYGSSWYFDVW	266	192	55
CARRLDYW	189	228	170
CMRYGSSWYFDVW	266	26	299
CTTVRYW	125	301	170
CARPYDYW	28	301	299
CAQMRGFAYW	125	464	40
CARFDYW	413	228	96
CMRYGSSWYFDVW	125	192	449
CARDGGYWYFDVW	189	301	299
CARRYGSSWYFDVW	125	464	219
CTTLRYW	266	112	449
CAKNWDYW	189	464	219
CARDYDYWYFDVW	413	192	299
CARROYGSSWYFDVW	24	918	52
CARYGPYFDYW	62	31	933
CARIYYGSSWYFDVW	266	464	299
CARDEFAYW	760	79	299
CARHYGSSWYFDVW	760	84	299
CARDWDYWYFDVW	266	464	449
CARGDGFYW	97	192	933
CARHYGSSWYFDVW	266	63	933
CARGYYW	760	57	449
CAKGDYGSSWFAYW	266	130	933
CARLYYGSSWYFDVW	97	301	933
CTRGYFDYW	760	158	449
CARGDGYFDYW	760	464	219
CARGGDYW	760	464	219
CARGAMDYW	189	464	933
CARYDGYFDYW	413	301	933
CTVYGGSTWFAYW	413	301	933
CARRDYW	760	464	449
CARSYFDYW	760	464	449
CAREGDYDWDYFDVW	760	918	52
CARDYYGSSGYFDVW	17	918	933
CTRVAYW	760	192	933
CTRWDYW	760	192	933
CARYAMDYW	760	228	933
CARDYYGSSFDYW	760	918	299
CARYSNYFDYW	760	918	299
CANYGSSWYFDVW	760	301	933
CARSLDYW	760	301	933
CASELGGFAYW	760	301	933
CASPNWDWYFDVW	760	301	933
CAREDYW	189	918	933
CARDGFAYW	266	918	933
CARKLDYW	760	464	933
CARNWDYAMDYW	760	464	933
CARYYYGSSWYFDVW	760	464	933

	Pool 1	Pool 2	Pool 3
CARSYFDYW	760	464	449
CAREGDYDWDYFDVW	760	918	52
CARDYYGSSGYFDVW	17	918	933
CTRVAYW	760	192	933
CTRWDYW	760	192	933
CARYAMDYW	760	192	933
CARDYYGSSFDYW	760	918	299
CARYSNYFDYW	760	918	299
CANYGSSWYFDVW	760	301	933
CARSLDYW	760	301	933
CASELGGFAYW	760	301	933
CASPNWDWYFDVW	760	301	933
CAREDYW	189	918	933
CARDGFAYW	266	918	933
CARKLDYW	760	464	933
CARNWDYAMDYW	760	464	933
CARYYYGSSWYFDVW	760	464	933
CARSGTDYW	413	918	933
CARFDYW	760	918	933
CARDGSSWYFDVW	760	918	933
CARDYFDYW	760	918	933
CARESNYFDYW	760	918	933
CARDWYFDVW	760	918	933
CARSYYAMDYW	760	918	933
CARYGNYAMDYW	760	918	933
CARYSNYAMDYW	760	918	933
CATGFAYW	760	918	933
CTGLYFDYW	760	918	933
CTYYGSSDFYW	760	918	933



B

	Pool 1	Pool 2	Pool 3
CQNGHSFPLTF	1	46	1
CQWSSYPFTF	9	3	24
CQWSSYPLTF	13	5	2
CQWSSNPFTF	2	19	7
CQQSNEDPRTF	10	20	9
CQQNTLPWTF	22	7	11
CQQNSYPLTF	32	8	3
CQHYSTPLTF	6	11	16
CFQGSHPYTF	16	12	21
CQHYSTPYTF	23	15	10
CFQGSHPWTF	17	15	17
CQQNSYPYTF	42	17	12
CLQYDEFYTF	38	18	17
CQQSNEDPYTF	13	22	19
CLQYASSPYTF	21	21	20
CLQHGESPYTF	19	14	23
CLQYDNLWTF	18	23	40
CSQSTHVPYTF	23	24	21
CQWNYPLTF	94	12	25
CSQSTHVPWTF	20	25	28
CQYSSYPLTF	25	8	33
CQWSSNPLTF	59	28	27
CQHYSTPWTF	34	29	26
CQWSSYPFTF	26	32	29
CQQNTLPYTF	26	34	46
CFQGSHPVPLTF	4	35	33
CQHYSTPRTF	71	31	33
CLQYDNLVTF	69	37	30
CQQNSWPFTF	6	171	39
CWQGTHPWTF	78	39	33
CMQHLEYPYTF	42	40	8
CQQGSYPLTF	94	41	41
CQQNTLPRTF	33	42	61
CQWSSNPYTF	86	38	43
CQWSSYPYTF	36	43	133
CQQSKEVPRTF	5	45	155
CQNGHSFPYTF	15	88	47
CLQSDNMLTF	46	46	42
CMQHLEYPFTF	26	46	56
CQQSNEDPWTF	80	46	13
CQQSNEDPFTF	26	128	51
CQYSKLPWTF	106	4	51
CQQSKEVPWTF	98	50	3
CQHHYGTPLTF	39	66	54
CQYSGYPLTF	98	43	54
CLQHGESPFTF	45	30	79
CWQGTHPRTF	39	60	56
CHQYLSWTF	46	53	69
CQHHYGTPYTF	56	54	61
CLQYASSPWTF	49	64	59
CWQGTHPYTF	49	70	32
CQHFVGTPTYTF	48	60	63
CHQYLSYTF	34	58	80
CQHYSTPFTF	65	59	47
CFQGSHPFTF	53	51	82
CQNDYSYPLTF	53	77	65
CQQNSWPLTF	63	60	37
CWQGTHPQTF	56	66	47
CQHFVGTPWTF	41	66	101
CFQGSYPLTF	60	6	99
CSQSTHVPPLTF	75	1	70
CQYSSYPWTF	63	98	14
CQDYSSPYTF	129	64	71

	Pool 1	Pool 2	Pool 3
CLQSDNLPLTF	67	70	67
CMQHLEYPPLTF	49	70	72
CQYNSYPFTF	129	70	47
CLQSDNMPYTF	65	75	72
CQYSSYPYTF	121	70	72
CQWSSNPFTF	67	77	51
CSQSTHVPFTF	56	96	76
CQHSRELPLTF	223	75	76
CLQYASYPRTF	49	81	128
CQQRSSYPLTF	71	98	43
CFQGSHPRTF	26	83	101
CQQHNEYPWTF	60	83	101
CQQGSYPWTF	74	33	82
CLQYDEFPLTF	106	81	87
CQQHNEYPPLTF	86	86	87
CQQHNEYPYTF	110	83	87
CQYSSYPLTF	53	94	149
CQNGHSFPFTF	10	216	92
CQQNTLPFTF	85	36	95
CQYSSYPYTF	98	96	72
CQHFVGTPLTF	117	10	95
CQQNSWPYTF	98	98	31
CLQYASSPFTF	86	87	114
CQQNSWPHFTF	157	88	99
CAQNLPLWTF	91	54	108
CQHHYGTPRTF	91	2	142
CQKSKEVPYTF	80	111	101
CLQYASSPLTF	121	103	85
CQQSSIPRTF	97	201	87
CSQSTHVPFTF	86	123	108
CQYSKLPYTF	157	106	92
CFQGSHPFTF	98	112	87
CLQSDNLPYTF	98	107	76
CQHFVGTPTF	98	77	219
CQQSNEDPLTF	98	123	65
CWQGTHPPLTF	80	161	112
CGQSYSYPYTF	108	112	92
CQHYSTPFTF	121	112	82
CQNDHSYPYTF	75	154	114
CHQYLSRTF	94	119	125
CAQNLPLYTF	114	119	117
CQHHYGTPFTF	117	123	114
CSQSTHVPRTF	109	123	125
CQQRSSYPFTF	12	143	118
CHQRSSWTF	183	128	5
CQHFVGTPTF	77	128	142
CQHHYGTPWTF	114	132	106
CAQYSSYPLTF	213	119	121
CQDYSSPLTF	137	98	121
CWQGTHPFTF	142	57	121
CHQYLSSLTF	157	132	112
CQNLSTPYTF	60	132	155
CQYSSYPFTF	321	102	125
CAQNLPLWTF	242	135	118
CQYSSYPFTF	171	135	45
CQHFVGTPTF	121	93	178
CQHFVSTPWTF	121	171	59
CQQSSIPPLTF	121	123	128
CQNDHSYPLTF	142	107	128
CQNLSTPFTF	91	185	128
CWQGTHPHTF	86	139	149
CQHSRELPLYTF	154	135	133
CQHFVSTPYTF	127	178	95

B

	Pool 1	Pool 2	Pool 3
CQNVLSTPWTF	149	143	106
CQQYWSTPYTF	200	143	56
CLQSDNMPFTF	149	143	135
CQQLVEYPYTF	265	139	135
CQQNNEPWF	129	135	135
CQQDYSSPWTF	129	147	118
CQQWSSFPFTF	129	142	465
CVQYAQFPYTF	129	103	142
CLQYASYPYTF	167	147	135
CQQGNTLPLTF	117	147	206
CQHGYGTPPTF	183	117	142
CQQGQSYPTF	137	60	142
CQOWSGYPFTF	149	128	142
CLQYDNLRTF	137	107	162
CQQWSSNPWF	137	219	101
CQNGHSFPFTF	69	238	149
CLQSDNLPFTF	142	154	135
CQNDYSYPTF	142	191	149
CQQSNEPPTF	142	161	155
CGQSYSPFTF	213	161	121
CLQHWNYPLTF	157	161	155
CQQNNEPRTF	149	147	219
CLQYDEFPTF	80	171	162
CLQYDEFPWTF	171	88	162
CHQRSSYPWF	129	167	174
CQQGSSIPFTF	167	167	162
CQQNNEPPTF	171	167	135
CQHFVWTPRTF	154	216	108
CQNGHSFPRTF	3	219	171
CQQYWSTPLTF	265	171	149
CQQYWSTPWTF	183	171	111
CLQYASYPWF	157	185	162
CQQLVEYPFTF	157	26	219
CQHSWEIPYTF	223	151	174
CVQYAQFPWF	200	154	174
CFQSNLYPYTF	78	178	219
CQQRSSYPFTF	167	178	162
CQQRSSYPYTF	183	178	174
CSQSTHVPPWF	183	178	162
CLQYASSPRTF	183	178	178
CQQLYSTPYTF	242	112	178
CLQYASSPFTF	171	185	155
CQQNNEPLTF	321	185	149
CQNDYSYPYTF	167	94	219
CQQHLHIPYTF	223	191	85
CQQSNSWPRTF	149	191	193
CSQSTHVPPYTF	157	191	193
CQQFTSSPYTF	200	191	182
CQQSKEVPFTF	171	56	182
CLQTHQPWF	171	171	306
CLQHGESPWF	110	201	240
CLQYDNLFTF	110	201	219
CLQVTHVPYTF	183	151	188
CQQYWSTPFTF	361	191	188
CFQSNLYPLTF	183	268	188
CHQRSSYTF	183	27	342
CQQNNEPFTF	183	205	68
CLQYDNLFTF	80	205	240
CQQGNTLPFTF	183	205	206
CQQWSSNPITF	223	205	155
CQQYHSYPLTF	42	205	240
CQQYYSYPRTF	110	205	193
CQHSRELPWF	293	119	193

	Pool 1	Pool 2	Pool 3
CQHSWEIPLTF	242	88	193
CQQLYSTPLTF	293	185	193
CLQYDNLTYTF	200	107	219
CQHSRELPFTF	200	219	193
CQQWSSFPYTF	200	154	465
CAGNLELPLTF	293	201	206
CLQYDNLWTF	213	77	206
CQQWSSNPRTF	559	88	206
CQNGHSFPWF	8	219	249
CQNVLSTPRTF	171	219	374
CQQDYSSPFTF	200	219	219
CAQNLPLRTF	213	268	206
CQHSRELPFTF	361	112	219
CQQLVEYPRTF	183	345	219
CQQYSKLPFTF	157	268	219
CSQSTHVPTF	129	256	219
CHQWSSYPLTF	265	228	37
CKQSYNLWTF	321	228	6
CQQGSSIPYTF	200	228	306
CVQYAQFPRTF	242	228	219
CHQWSSYPYTF	223	320	193
CLQYDEFRTF	223	320	206
CQQLVEYPLTF	223	191	282
CQQYSSYPWF	223	205	374
CHQRSSYPYTF	183	238	249
CKQSYNLTF	265	238	219
CLQYDNLRTF	200	238	282
CQHFVSTPRTF	183	238	342
CQNVLSTPPTF	361	238	128
CQQDYSSPFTF	223	238	262
CQQLYSTPWTF	223	238	262
CQSYSAPLTF	293	238	240
CQQYSKLPRTF	321	238	81
CKQSYNLTYTF	242	268	162
CQQGQSYPTF	242	117	282
CQQHNEYPTF	242	320	63
CQQWSNYPFTF	242	238	625
CQQYYSYPTF	242	219	282
CGQSYSPFTF	293	51	249
CHQWSSYPTF	265	167	249
CQHSWEIPWF	223	268	249
CHQYHRSPLTF	183	256	342
CQQDYSSPRTF	223	256	342
CQQYNSYPRTF	361	256	182
CQQYWSTPPTF	293	256	240
CLQTHQPYTF	559	268	193
CQQGNTLWTF	559	268	219
CQQSNSWPWF	427	268	188
CQQYNNYPLTF	559	268	219
CQQYSGYPYTF	242	268	282
CHQRSSYPFTF	213	381	262
CKQAYDVYTF	293	268	262
CLQYDNLFTF	321	154	262
CQSRKVPWF	321	205	262
CSQSTHVPPFTF	361	228	262
CSQSTHVPLTF	293	216	262
CFQSGYPTF	265	238	374
CKQSYNLFTF	265	105	342
CLQYDEFPTF	265	219	465
CQNVLSTPLTF	265	256	306
CQQFTSSPWTF	265	289	262
CQQGNTLYTF	265	238	374
CQQWNSYPLTF	265	421	142

B

	Pool 1	Pool 2	Pool 3
CQQYHSYPPTF	265	345	219
CAQNLELPPTF	559	289	219
CHQWSSYPPTF	293	289	249
CHQWSSYPWTF	321	289	240
CHQYHRSPPTF	223	289	465
CHQYHRSPYTF	265	289	342
CLOHWNYPYTF	171	289	625
CQQDYSSYTF	242	289	465
CQOFTSSPPTF	559	289	219
CQQSKEVPPTF	559	289	15
CQQYNSYPWTF	427	289	182
CQQYSSYPRTF	242	289	306
CSQSTHVWTF	321	289	249
CLQYASYPPTF	361	69	282
CQQWSGYPYTF	242	381	282
CQQYSKLPPTF	293	185	282
CFQGSHPVPTF	293	289	465
CHQYHRSPWTF	293	191	374
CLQHSYLPYTF	293	421	219
CQQGNTLPPWTF	293	289	625
CQQWNYPYTF	293	238	625
CQQYSKLPPTF	293	238	306
CLOGTHOPPTF	293	320	625
CQHFWDTPRTF	213	320	306
CQHFWSPTPTF	265	320	306
CQQRSSYPWTF	427	320	262
CQQWSSSPPTF	265	320	374
CQQWSSYPWTF	321	320	282
CQQYSSYPPTF	223	320	625
CVQYAQFPPTF	265	320	374
CLOHGESPLTF	427	320	306
CLOHWNYPPTF	114	381	306
CLQRNAYPLTF	361	289	306
CQHSWEIPPTF	559	228	306
CQOFTSSPSTF	200	504	306
CQQHYSSPLTF	265	646	306
CQQNNEPPTF	427	256	306
CQQRKVPYTF	427	289	306
CQQWSGYPLTF	427	320	306
CQQWTYPLTF	321	205	306
CSQSTHVLTF	559	289	306
CLQVTHVPWTF	321	345	282
CQHFWSPLTF	321	421	206
CQNDHSYPPTF	321	345	240
CQQYSGYPPTF	321	421	262
CQQYWSTPRTF	321	345	262
CVQYAQFPPTF	321	381	188
CHQYHRSPPTF	242	345	374
CLQYASYPLTF	427	345	306
CLQYDEFPPYTF	171	345	342
CQQGNTLRTF	361	345	171
CQQSKEVPLTF	427	345	162
CQQWSSYPITF	242	345	465
CSQSTHVYTF	265	345	374
CWQGTHFPF	321	345	625
CLQYDELYTF	242	504	342
CQQLYSTPRTF	427	289	342
CQQSNSWPQYTF	361	345	342
CQQWSSYPRTF	265	381	342
CQQYNSYPLTF	559	345	342
CAQNLELPPTF	427	381	206
CHQRSSFTF	200	381	374
CQQYSGYPWTF	321	381	374

	Pool 1	Pool 2	Pool 3
CQQYSTYPYTF	559	381	306
CHQRSSYPLTF	361	504	342
CHQYLSSTF	361	381	465
CKQAYDVPLTF	361	345	625
CLQYEFPLTF	361	268	465
CLQYEFYPTF	361	345	465
CQQGSSIPRTF	361	381	625
CQQHLHIPWTF	361	646	262
COQHYSTPTF	361	289	374
CQQSNSWPQLTF	361	421	342
CQQYSSSPLTF	361	191	374
CVQGTHFPYTF	361	289	465
CGQSYSYPTF	559	320	374
CHQRSSYPCTF	321	646	374
CHQWSSYPPTF	183	421	374
CHQWSSYRTF	559	228	374
CHQYLSLTF	427	289	374
CLQRNAYPYTF	242	421	374
CLQVTHVPPTF	361	504	374
CQNGHSFPF	223	646	374
CQQGQSYPTF	427	289	374
CQQRSSYPRTF	559	345	374
CQQWNYPLYTF	427	345	374
CQQWSSNPPYTF	427	381	374
CQQYNSYPHTF	321	646	374
CQQYYSYTF	559	345	374
CSQSTHIPWTF	293	646	374
CGQSYSYPPTF	223	421	625
CHQYLSSTF	559	421	262
CLOGTHQPPTF	427	421	374
CLQVTHVPPTF	559	421	374
CQHFWSPTPTF	321	421	625
CQHSWEIPPTF	427	421	306
CQQSNSWPQTF	559	421	282
CQQYHSYPRTF	427	421	249
CQQYHSYPYTF	427	421	282
CAQNLELPTF	427	421	625
CFQGSHPHTF	427	268	625
CHQRSSYPF	427	421	625
CHQYLSYTF	427	421	625
CKQSYNLRTF	427	268	625
CLQYASSLTF	427	421	625
CLQYDNLTLTF	427	646	306
CQHSRELPRTF	427	504	193
CQQGSSSPYTF	427	421	625
CQQWNNYPLTF	427	646	374
CQQWSSSPLTF	427	646	374
CQQWSSYPITF	427	320	625
CQQYNKLPWTF	427	345	625
CQQYNSYLTf	427	345	465
CQQYSSYTF	427	381	625
CQQYSSYPTF	427	345	625
CQQYSSYPF	427	421	625
CQQYSSYPPTF	427	421	465
CSQSTHVFPF	427	646	374
CVQGTHFPRTF	427	504	342
CWQGTHFRTF	427	345	465
CKQSYNLPTF	559	268	465
CLQHSYLPPTF	559	421	465
CQHSWEIPRTF	559	421	465
CQQGNTLPLTF	559	421	465
CQQLYSTPPTF	361	504	465
CQQSNSWPF	361	504	465

B

	Pool 1	Pool 2	Pool 3
CQQYHSYPWTF	183	646	465
CQQYSGYPLTF	293	646	465
CQQYSKLWTF	427	504	465
CQQYSSYRTF	157	504	465
CWQGTHFPTF	321	646	465
CCQGSYVPLTF	559	504	374
CFQSNYLPFTF	559	504	374
CHQWSSYHTF	559	504	465
CLQYDNLWTF	559	504	465
CQHFWNTPYTF	559	504	249
CQQFTSSPLTF	559	504	374
CQQSYSAPFTF	559	504	374
CQWSSDPFTF	223	504	625
CQWSSNPPMYTF	321	504	625
CQWSSNPQYTF	427	504	625
CQQNSYPF	361	504	625
CQQNSYPLYTF	559	504	465
CVQGTHFPHTF	559	504	374
CVQGTHFPLTF	559	504	306
CGQSYSYPRTF	559	504	625
CHQYHRSPRTF	559	289	625
CHQYLSPTF	559	646	465
CLQHGESPF	559	504	625
CLQYDNLPTF	559	646	465
CLQYDSLWTF	559	421	625
CMQOLEYPYTF	559	646	465
CQNDHSYPWTF	559	646	465
CQHYSTPF	559	646	282
CQHYSTWTF	559	646	374
CQQRSSYPPLTF	559	646	465
CQQRKVPSTF	559	646	374
CQWSSNPPWTF	559	504	625
CQWSSYPLTF	559	646	374
CQQNSYPLAF	559	646	342
CQQNSYTF	559	646	465
CQQYSGYPRTF	559	646	342
CQQYSSYPLTF	559	421	625
CQQYTSYPLTF	559	268	625
CQQYWSTPPWTF	559	646	465
CSQSTHVPHTF	559	320	625
CVQYAQFPPTF	559	646	465
CFQGSYHGPWTF	559	646	625
CFQGSYVPTF	559	646	625
CHQRSSYPTF	559	646	625
CLQGTYYPRTF	559	646	625
CLQYASSPYMYTF	427	646	625
CLQYDEFPPF	223	646	625
CLQYDEFPPPLTF	427	646	625
CMQHLECPYTF	559	646	625
CQHHYGTWTF	427	646	625
CQNDHSYPRTF	559	646	625
CQHYSTPHTF	559	646	625
CQOSIEDPYTF	559	646	625
CQOSNEDPPWTF	361	646	625
CQQYNTYPLTF	427	646	625
CQQYWSTALTF	559	646	625
CQQYYSYRTF	361	646	625
CSQSTHVPTWTF	361	646	625

Rankings of CDR3 sequences shared by all three mouse pools were uniform in both IgH (A) and Igκ (B). The most abundant CDR3 sequence is ranked as 1. Dark red indicates higher rank moving to blue, of lower rank.

## **Appendix A.5 Alignment of top IgH gene segment combination**

(See following page.)

Full CDR3 nucleotide alignment of IgH gene combination examined in Figure 7 (IGHV1-26, IGHD1-1, IGHJ1).



## **Appendix A.6 V-gene segment heat maps**

(See following page.)

(A) VH- and (B) V $\kappa$ -gene segment usage in animals within ground (G) and flight (F) treatment groups are presented as percent of repertoire and rank. V-gene segments are listed by rank order (most frequent to least frequent). Dark red indicates higher percent of repertoire or rank moving to blue, lower percent of repertoire or rank. V-gene segments with identical ranks are displayed as ties.

A

	G1	G2	G3	F1	F2	F3	Avg G	Avg F	Avg T		G1	G2	G3	F1	F2	F3
V1-53*01	12.78	8.74	23.71	10.15	4.99	8.83	15.08	7.99	11.53	V1-53*01	1	2	1	2	2	1
V9-3*01	10.07	22.08	5.81	2.79	3.10	7.88	12.65	4.59	8.62	V9-3*01	2	1	2	8	6	2
V1-26*01	3.87	3.20	3.88	9.70	3.32	5.63	3.65	6.22	4.93	V1-26*01	4	5	4	3	5	4
V3-6*01	3.24	5.20	4.58	11.29	1.86	2.07	4.34	5.07	4.71	V3-6*01	6	3	3	1	14	12
V1-78*01	0.72	0.73	0.90	0.69	16.63	1.75	0.78	6.36	3.57	V1-78*01	36	31	28	34	1	17
V6-3*01	3.40	2.35	2.13	1.49	4.30	7.48	2.63	4.43	3.53	V6-3*01	5	9	11	13	3	3
V5-4*01	0.96	1.10	1.38	5.45	1.30	1.26	1.15	2.67	1.91	V5-4*01	30	21	16	5	20	19
V1-15*01	0.25	1.04	0.90	6.46	1.09	0.68	0.73	2.75	1.74	V1-15*01	64	23	27	4	29	36
V1-19*01	5.93	0.42	0.71	0.42	0.81	0.83	2.35	0.69	1.52	V1-19*01	3	52	37	47	36	31
V1-55*01	2.53	3.64	2.71	2.96	1.84	4.76	2.96	3.19	3.07	V1-55*01	10	4	7	7	15	5
V1-82*01	1.63	1.63	2.83	2.44	2.68	2.83	2.03	2.72	2.37	V1-82*01	17	13	6	10	7	10
V1-9*01	1.97	1.83	2.44	3.56	1.24	1.21	2.08	2.00	2.04	V1-9*01	13	10	9	6	23	24
V1-80*01	1.55	1.72	2.16	2.42	1.81	1.99	1.81	2.07	1.94	V1-80*01	20	11	10	11	16	15
V1-72*01	1.83	1.28	1.95	1.45	2.43	2.02	1.69	1.96	1.83	V1-72*01	15	18	13	16	9	13
V2-2*01	1.13	1.05	2.92	1.36	1.25	3.18	1.70	1.93	1.82	V2-2*01	26	22	5	17	22	9
V1-64*01	1.86	2.59	1.26	1.04	2.18	1.68	1.90	1.63	1.77	V1-64*01	14	8	18	24	11	18
V1-81*01	1.17	0.69	0.84	1.07	1.19	3.74	0.90	2.00	1.45	V1-81*01	25	32	31	22	25	6
V1-74*01	2.97	0.76	0.81	0.96	0.53	2.57	1.51	1.35	1.43	V1-74*01	7	29	32	26	53	11
V9-1*01	1.57	2.75	0.77	0.19	2.28	0.64	1.70	1.03	1.37	V9-1*01	18	7	34	67	10	39
V1-62-2*01	2.58	0.44	0.23	0.19	0.68	3.31	1.09	1.39	1.24	V1-62-2*01	8	50	65	65	42	7
V1-71*01	2.58	0.44	0.23	0.19	0.68	3.31	1.08	1.39	1.24	V1-71*01	8	51	66	65	42	7
V5-6*01	1.66	0.85	1.41	0.96	1.27	1.25	1.31	1.16	1.23	V5-6*01	16	27	15	27	21	20
V11-2*01	1.57	1.49	0.77	1.07	1.31	1.13	1.28	1.17	1.22	V11-2*01	19	15	34	23	19	27
V2-9-1*01	0.43	1.46	0.96	0.47	1.94	2.01	0.95	1.47	1.21	V2-9-1*01	48	16	26	45	13	14
V14-4*01	1.25	0.54	0.96	1.56	0.78	1.91	0.92	1.42	1.17	V14-4*01	23	40	25	12	37	16
V1-52*01	1.19	1.19	1.17	1.13	1.19	1.08	1.19	1.13	1.16	V1-52*01	24	20	19	20	24	29
V1-76*01	0.48	0.40	1.12	1.49	2.74	0.55	0.66	1.59	1.13	V1-76*01	46	55	20	14	8	45
V2-3*01	2.22	0.39	0.98	0.57	1.64	0.80	1.20	1.00	1.10	V2-3*01	11	56	24	37	17	33
V4-1*01	0.84	0.51	0.34	0.36	4.10	0.30	0.56	1.59	1.07	V4-1*01	32	45	57	51	4	56
V14-2*01	0.81	0.51	2.61	1.20	0.67	0.60	1.31	0.83	1.07	V14-2*01	33	44	8	18	45	40
V1-75*01	0.86	0.79	1.63	0.93	0.87	1.21	1.09	1.00	1.05	V1-75*01	31	28	14	28	33	22
V1-22*01	0.32	1.58	0.65	2.51	0.68	0.52	0.85	1.24	1.04	V1-22*01	58	14	42	9	41	47
V5-17*01	0.56	1.33	1.98	0.62	1.05	0.67	1.29	0.78	1.04	V5-17*01	43	17	12	36	32	37
V9-2*01	1.38	3.00	0.71	0.07	0.15	0.41	1.70	0.21	0.95	V9-2*01	21	6	36	83	73	49
V1-39*01	0.55	0.94	0.60	1.45	1.08	0.94	0.70	1.16	0.93	V1-39*01	44	24	43	15	30	30
V1-50*01	0.99	1.24	0.78	0.81	0.87	0.75	1.00	0.81	0.91	V1-50*01	29	19	33	32	34	34
V8-8*01	2.12	0.47	0.57	1.11	0.59	0.26	1.05	0.66	0.85	V8-8*01	12	48	45	21	50	61
V1-69*01	1.00	0.62	0.87	0.63	0.66	1.21	0.83	0.83	0.83	V1-69*01	28	34	29	35	46	23
V6-6*01	0.68	0.59	1.12	0.54	1.17	0.71	0.80	0.81	0.80	V6-6*01	40	36	20	38	27	35
V5-16*01	1.12	0.52	0.42	0.47	1.06	1.19	0.69	0.91	0.80	V5-16*01	27	43	53	43	31	25
V1-7*01	0.69	0.28	0.53	0.84	2.09	0.26	0.50	1.07	0.78	V1-7*01	39	67	46	31	12	60
V1-18*01	0.33	0.76	1.34	0.97	0.71	0.53	0.81	0.74	0.78	V1-18*01	56	30	17	25	39	46
V1-42*01	0.71	0.54	1.00	0.91	0.65	0.81	0.75	0.79	0.77	V1-42*01	38	41	23	29	47	32
V8-12*01	1.31	0.59	1.00	0.52	0.46	0.57	0.97	0.52	0.74	V8-12*01	22	37	22	39	57	44
V5-9-1*02	0.28	0.66	0.37	0.28	1.45	1.16	0.44	0.96	0.70	V5-9-1*02	62	33	55	58	18	26
V1-59*01	0.64	0.92	0.87	0.35	0.68	0.57	0.81	0.54	0.67	V1-59*01	41	25	30	54	44	43
V1-61*01	0.77	0.59	0.67	1.18	0.44	0.32	0.67	0.65	0.66	V1-61*01	34	37	41	19	59	54
V2-9*01	0.35	0.49	0.36	0.45	1.16	1.13	0.40	0.91	0.66	V2-9*01	54	46	56	46	28	28
V7-3*01	0.76	0.59	0.43	0.35	0.78	0.65	0.59	0.59	0.59	V7-3*01	35	39	51	55	38	38
V2-5*01	0.18	1.64	0.17	0.49	0.46	0.27	0.67	0.40	0.53	V2-5*01	68	12	71	42	58	59
V14-3*01	0.31	0.53	0.50	0.85	0.59	0.36	0.45	0.60	0.52	V14-3*01	59	42	49	30	51	50
V3-1*01	0.52	0.41	0.51	0.76	0.50	0.35	0.48	0.54	0.51	V3-1*01	45	54	48	33	54	51
V10-1*01	0.33	0.60	0.53	0.38	0.61	0.50	0.49	0.50	0.49	V10-1*01	57	35	47	49	48	48
V9-4*01	0.19	0.14	0.22	0.14	0.69	1.23	0.18	0.69	0.43	V9-4*01	66	79	67	73	40	21
V1-54*01	0.41	0.39	0.29	0.27	0.82	0.31	0.36	0.47	0.41	V1-54*01	49	57	60	60	35	55
V10-3*01	0.35	0.27	0.70	0.52	0.21	0.22	0.44	0.32	0.38	V10-3*01	55	68	39	41	65	64
V1-77*01	0.36	0.31	0.18	0.28	0.48	0.59	0.28	0.45	0.37	V1-77*01	53	65	69	59	56	41
V1-34*01	0.71	0.33	0.39	0.34	0.19	0.21	0.48	0.25	0.36	V1-34*01	37	61	54	56	67	65
V14-1*01	0.59	0.48	0.57	0.20	0.16	0.14	0.55	0.17	0.36	V14-1*01	42	47	44	64	72	71
V5-12*01	0.12	0.31	0.14	0.09	1.19	0.22	0.19	0.50	0.34	V5-12*01	74	64	74	78	25	63
V1-12*01	0.29	0.16	0.27	0.39	0.49	0.29	0.24	0.39	0.32	V1-12*01	60	73	61	48	55	57
V3-8*01	0.46	0.39	0.21	0.37	0.28	0.16	0.35	0.27	0.31	V3-8*01	47	57	68	50	61	70
V1-47*01	0.14	0.35	0.13	0.26	0.55	0.34	0.21	0.38	0.29	V1-47*01	72	60	76	61	52	52



A

	G1	G2	G3	F1	F2	F3	Avg G	Avg F	Avg T		G1	G2	G3	F1	F2	F3
V1-85*01	0.11	0.24	0.49	0.23	0.09	0.58	0.28	0.30	0.29	V1-85*01	77	70	50	62	76	42
V1-66*01	0.15	0.18	0.25	0.52	0.23	0.27	0.19	0.34	0.27	V1-66*01	71	71	63	40	63	58
V1-20*01	0.18	0.32	0.43	0.09	0.17	0.33	0.31	0.20	0.26	V1-20*01	67	62	51	77	71	53
V1-5*01	0.12	0.27	0.23	0.47	0.18	0.19	0.21	0.28	0.25	V1-5*01	74	69	64	44	70	66
V1-36*01	0.11	0.92	0.07	0.16	0.07	0.08	0.37	0.10	0.23	V1-36*01	78	26	84	69	80	80
V1-31*01	0.41	0.17	0.68	0.02	0.03	0.02	0.42	0.02	0.22	V1-31*01	50	72	40	91	94	86
V1-58*01	0.10	0.47	0.30	0.21	0.12	0.12	0.29	0.15	0.22	V1-58*01	79	49	58	63	75	75
V1-4*01	0.37	0.29	0.17	0.08	0.19	0.18	0.27	0.15	0.21	V1-4*01	52	66	72	80	68	68
V1-63*01	0.09	0.08	0.71	0.09	0.08	0.19	0.29	0.12	0.21	V1-63*01	81	86	38	79	78	67
V5-9*01	0.13	0.15	0.18	0.15	0.40	0.14	0.15	0.23	0.19	V5-9*01	73	76	70	70	60	73
V1-84*01	0.29	0.16	0.26	0.07	0.20	0.11	0.23	0.13	0.18	V1-84*01	61	74	62	82	66	76
V1-62-3*01	0.40	0.31	0.29	0.04	0.02	0.02	0.33	0.03	0.18	V1-62-3*01	51	63	59	88	100	89
V1-11*01	0.09	0.36	0.05	0.04	0.15	0.23	0.17	0.14	0.15	V1-11*01	83	59	85	85	74	62
V13-2*01	0.27	0.11	0.10	0.12	0.22	0.10	0.16	0.15	0.15	V13-2*01	63	83	80	76	64	77
V8-5*01	0.23	0.12	0.10	0.12	0.18	0.13	0.15	0.15	0.15	V8-5*01	65	81	78	75	69	74
V2-6*01	0.15	0.14	0.14	0.35	0.06	0.02	0.14	0.14	0.14	V2-6*01	69	78	73	52	83	90
V2-6-8*01	0.15	0.14	0.13	0.35	0.06	0.02	0.14	0.14	0.14	V2-6-8*01	70	80	75	52	83	90
V3-5*01	0.12	0.15	0.08	0.08	0.25	0.14	0.12	0.16	0.14	V3-5*01	74	77	83	81	62	72
V3-4*01	0.03	0.04	0.02	0.04	0.61	0.04	0.03	0.23	0.13	V3-4*01	94	94	99	86	49	83
V12-3*01	0.08	0.11	0.13	0.15	0.09	0.18	0.11	0.14	0.12	V12-3*01	86	82	77	71	77	69
V2-4*01	0.09	0.41	0.04	0.07	0.04	0.08	0.18	0.06	0.12	V2-4*01	82	53	90	84	87	79
V5-15*01	0.08	0.06	0.10	0.13	0.07	0.10	0.08	0.10	0.09	V5-15*01	87	88	79	74	79	78
V1-37*01	0.02	0.08	0.05	0.33	0.04	0.02	0.05	0.13	0.09	V1-37*01	99	85	86	57	86	93
V1-23*01	0.04	0.09	0.10	0.16	0.04	0.02	0.07	0.08	0.07	V1-23*01	91	84	82	68	85	87
V1-56*01	0.08	0.16	0.10	0.02	0.02	0.01	0.11	0.02	0.06	V1-56*01	85	74	80	95	98	96
V5-2*01	0.01	0.05	0.04	0.15	0.03	0.05	0.03	0.07	0.05	V5-2*01	117	89	88	72	92	82
V8-2*01	0.10	0.04	0.01	0.02	0.06	0.03	0.05	0.04	0.05	V8-2*01	80	91	105	92	81	85
V8-11*01	0.09	0.05	0.05	0.04	0.02	0.01	0.06	0.02	0.04	V8-11*01	84	90	87	86	95	99
V1-49*01	0.03	0.06	0.03	0.01	0.03	0.01	0.04	0.02	0.03	V1-49*01	92	87	91	98	92	94
V1-43*01	0.05	0.01	0.03	0.02	0.01	0.03	0.03	0.02	0.03	V1-43*01	88	106	92	90	109	84
V7-4*01	0.01	0.02	0.02	0.01	0.03	0.05	0.02	0.03	0.03	V7-4*01	103	101	95	98	89	81
V1-67*01	0.03	0	0.02	0.02	0.06	0.02	0.02	0.03	0.02	V1-67*01	96	113	97	93	82	88
V15-2*01	0.03	0.02	0.03	0.03	0.02	0.01	0.03	0.02	0.02	V15-2*01	95	100	94	89	99	97
V1-8*01	0.04	0	0.02	0.02	0.03	0.02	0.02	0.02	0.02	V1-8*01	89	113	102	95	90	82
V3-3*01	0.01	0.02	0.02	0.02	0.04	0.00	0.02	0.02	0.02	V3-3*01	105	103	103	93	88	115
V8-9*01	0.01	0.04	0.01	0.01	0.01	0.00	0.02	0.01	0.02	V8-9*01	100	91	107	103	101	112
V1-21-1*01	0.03	0	0.04	0.00	0.01	0.00	0.02	0.01	0.02	V1-21-1*01	93	113	89	112	107	109
V6-7*01	0.01	0.03	0.02	0.01	0.01	0.01	0.02	0.01	0.01	V6-7*01	115	97	103	104	102	103
V2-7*01	0.01	0.02	0.01	0.01	0.03	0.01	0.01	0.01	0.01	V2-7*01	103	103	116	108	90	99
V1-17-1*01	0.01	0.04	0.02	0.00	0.00	0	0.02	0.00	0.01	V1-17-1*01	108	91	96	116	119	123
V7-2*01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	V7-2*01	110	107	110	97	103	98
V16-1*01	0.01	0.02	0.01	0.01	0.01	0.00	0.02	0.01	0.01	V16-1*01	115	99	105	101	104	110
V1-16*01	0.01	0.04	0.02	0.00	0	0	0.02	0.00	0.01	V1-16*01	108	94	100	122	124	123
V1-27*01	0.03	0	0.03	0.00	0.01	0.00	0.02	0.00	0.01	V1-27*01	96	113	93	118	109	114
V1-24*01	0.01	0.04	0.02	0.00	0	0	0.02	0.00	0.01	V1-24*01	105	96	101	122	124	123
V8-6*01	0.02	0.03	0.00	0.00	0.01	0.01	0.02	0.01	0.01	V8-6*01	98	98	124	125	106	102
V1-14*01	0.04	0.01	0.00	0.01	0.00	0.00	0.02	0.01	0.01	V1-14*01	90	110	125	105	114	106
V6-4*01	0.00	0.01	0.01	0.01	0.02	0.00	0.01	0.01	0.01	V6-4*01	119	109	115	101	96	106
V8-4*01	0.01	0.01	0.01	0.01	0.02	0.00	0.01	0.01	0.01	V8-4*01	112	111	118	98	97	110
V5-12-4*01	0.01	0.02	0.01	0.00	0.01	0.01	0.01	0.01	0.01	V5-12-4*01	101	102	118	111	109	103
V11-1*01	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.01	V11-1*01	105	108	109	115	114	115
V1-51*01	0.01	0	0.01	0.00	0.01	0.00	0.01	0.01	0.01	V1-51*01	101	113	112	112	104	105
V1-62-1*01	0.01	0.02	0.01	0.00	0	0	0.01	0.00	0.01	V1-62-1*01	111	105	108	122	124	123
V1-48*01	0.01	0	0.02	0.00	0.00	0.00	0.01	0.00	0.01	V1-48*01	114	113	98	116	114	112
V6-5*01	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.01	V6-5*01	124	112	118	108	113	95
V5-21*01	0.00	0	0.01	0.01	0.01	0.01	0.00	0.01	0.01	V5-21*01	124	113	113	105	107	99
V5-1*01	0.01	0	0.01	0.00	0.01	0.00	0.01	0.00	0.01	V5-1*01	112	113	110	112	112	115
V3-7*01	0.00	0	0.01	0.01	0.00	0.00	0.00	0.00	0.00	V3-7*01	124	113	118	105	118	115
V1-13*01	0.00	0	0.00	0.01	0	0.00	0.00	0.00	0.00	V1-13*01	122	113	126	108	124	106
V1-19-1*01	0.00	0	0.01	0.00	0.00	0.00	0.00	0.00	0.00	V1-19-1*01	119	113	113	125	121	119
V1-32*01	0.00	0	0.01	0.00	0.00	0.00	0.00	0.00	0.00	V1-32*01	119	113	117	120	121	119
V1-35*01	0.00	0	0.01	0.00	0.00	0.00	0.00	0.00	0.00	V1-35*01	118	113	122	120	121	119
V12-2*01	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	V12-2*01	122	113	123	118	114	123
V1-60*01	0	0	0.00	0	0.00	0.00	0.00	0.00	0.00	V1-60*01	127	113	126	127	119	122

B

	G1	G2	G3	F1	F2	F3	Avg G	Avg F	Avg T		G1	G2	G3	F1	F2	F3
V5-39*01	17.87	35.97	21.10	16.89	17.08	20.64	24.98	18.20	21.59	V5-39*01	1	1	1	1	1	1
V3-4*01	15.33	8.62	12.80	4.70	3.43	9.05	12.25	5.73	8.99	V3-4*01	2	2	2	5	5	2
V1-117*01	2.42	2.23	2.34	8.60	4.03	1.97	2.33	4.87	3.60	V1-117*01	6	7	7	2	3	10
V2-137*01	1.57	1.86	12.18	2.54	1.71	1.37	5.20	1.87	3.54	V2-137*01	16	9	3	8	13	19
V1-110*01	2.15	1.84	1.88	1.98	2.68	8.65	1.96	4.44	3.20	V1-110*01	9	10	11	14	8	3
V10-98*01	2.39	2.37	2.04	5.31	3.29	1.90	2.27	3.50	2.88	V10-98*01	8	6	10	4	6	12
V4-81*01	0.25	0.41	0.28	0.30	13.91	0.58	0.31	4.92	2.62	V4-81*01	55	49	55	54	2	37
V4-55*01	0.96	2.84	2.27	4.33	0.50	1.24	2.02	2.02	2.02	V4-55*01	24	3	8	6	44	21
V14-111*01	2.40	1.26	2.16	2.43	1.04	2.67	1.94	2.05	1.99	V14-111*01	7	13	9	11	24	6
V6-25*01	9.26	0.54	0.58	0.79	0.26	0.30	3.45	0.45	1.95	V6-25*01	3	38	35	29	57	54
V14-126-1*01	2.50	2.54	1.25	1.49	1.86	1.74	2.10	1.89	1.90	V14-126-1*01	5	5	18	19	12	16
V6-15*01	1.16	2.61	1.39	2.13	1.55	1.95	1.72	1.88	1.80	V6-15*01	20	4	17	13	16	11
V3-2*01	1.48	2.04	0.96	1.34	3.01	1.88	1.49	2.08	1.79	V3-2*01	19	8	21	20	7	13
V12-44*01	1.98	1.21	1.05	2.47	1.64	1.68	1.42	1.92	1.67	V12-44*01	10	15	20	9	15	17
V4-68*01	0.50	0.67	3.27	0.54	1.20	3.70	1.48	1.81	1.65	V4-68*01	40	33	4	37	21	4
V6-17*01	1.54	0.98	0.61	1.58	3.99	0.80	1.04	2.12	1.58	V6-17*01	18	20	32	17	4	31
V1-135*01	1.13	1.12	2.72	0.89	2.17	1.30	1.66	1.45	1.56	V1-135*01	21	16	5	26	10	20
V4-57-1*01	0.95	1.39	2.40	1.03	1.29	2.18	1.58	1.50	1.54	V4-57-1*01	25	12	6	23	17	9
V19-93*01	1.89	0.96	0.83	2.56	0.84	1.04	1.23	1.48	1.35	V19-93*01	11	21	24	7	32	26
V8-24*01	0.89	0.75	1.40	2.48	1.66	0.94	1.01	1.69	1.35	V8-24*01	26	28	16	10	14	27
V6-13*01	0.34	0.19	0.11	6.98	0.13	0.08	0.22	2.39	1.30	V6-13*01	51	65	67	3	74	75
V3-5*01	1.75	1.58	0.82	1.50	0.66	1.21	1.38	1.12	1.25	V3-5*01	14	11	26	18	39	22
V4-59*01	0.89	0.75	1.49	1.08	2.54	0.87	0.98	1.50	1.24	V4-59*01	31	27	14	22	9	29
V6-20*01	1.85	0.64	0.65	1.77	0.88	1.15	1.05	1.27	1.16	V6-20*01	12	35	29	15	29	23
V16-104*01	3.47	0.47	0.40	0.90	0.89	0.65	1.44	0.81	1.13	V16-104*01	4	43	43	25	28	34
V5-43*01	0.65	0.93	0.65	0.51	0.68	3.03	0.74	1.41	1.08	V5-43*01	32	22	30	38	38	5
V6-23*01	0.45	0.83	1.43	2.15	0.91	0.51	0.91	1.19	1.05	V6-23*01	44	25	15	12	27	38
V12-46*01	0.62	0.99	1.16	0.61	1.08	1.75	0.93	1.15	1.04	V12-46*01	33	19	19	34	22	15
V4-91*01	0.44	0.52	0.63	0.62	1.28	2.66	0.53	1.52	1.03	V4-91*01	45	41	31	33	18	7
V17-127*01	0.60	1.06	0.80	0.85	0.84	1.79	0.82	1.16	0.99	V17-127*01	34	18	27	27	31	14
V6-32*01	0.75	0.68	0.82	0.63	2.12	0.78	0.75	1.18	0.96	V6-32*01	28	32	25	32	11	33
V9-120*01	1.55	0.69	0.90	0.80	1.03	0.82	1.05	0.88	0.96	V9-120*01	17	31	22	28	25	30
V4-53*01	0.48	0.92	0.35	0.26	1.04	2.58	0.58	1.29	0.94	V4-53*01	42	23	49	58	23	8
V3-12-1*01	1.60	0.73	0.56	0.92	0.54	1.08	0.96	0.85	0.90	V3-12-1*01	15	29	34	24	41	24
V9-124*01	0.23	0.54	1.62	1.25	0.75	0.92	0.80	0.97	0.89	V9-124*01	58	39	13	21	36	28
V17-121*01	0.45	1.09	0.34	0.41	1.26	1.47	0.63	1.05	0.84	V17-121*01	43	17	50	47	20	18
V13-85*01	1.07	0.42	1.69	0.42	0.53	0.47	1.06	0.47	0.76	V13-85*01	22	48	12	45	43	42
V3-10*01	0.71	0.59	0.83	0.46	1.26	0.51	0.71	0.74	0.73	V3-10*01	29	37	23	43	19	39
V10-94*01	0.34	0.61	0.49	1.70	0.53	0.40	0.48	0.88	0.68	V10-94*01	52	36	40	16	42	46
V8-30*01	0.52	0.65	0.58	0.70	0.82	0.30	0.58	0.61	0.59	V8-30*01	38	34	33	31	34	53
V8-19*01	0.78	0.70	0.40	0.28	0.71	0.59	0.63	0.53	0.58	V8-19*01	27	30	44	57	37	36
V2-109*01	0.51	0.54	0.55	0.50	0.78	0.34	0.53	0.54	0.54	V2-109*01	39	40	37	39	35	49
V3-7*01	0.23	0.38	0.73	0.26	0.83	0.78	0.44	0.62	0.53	V3-7*01	57	51	28	58	33	32
V4-72*01	0.42	1.23	0.41	0.48	0.31	0.31	0.69	0.37	0.53	V4-72*01	47	14	42	41	54	50
V8-27*01	0.53	0.44	0.51	0.32	0.86	0.49	0.49	0.56	0.53	V8-27*01	37	45	38	52	30	41
V13-84*01	0.70	0.90	0.21	0.40	0.47	0.42	0.60	0.43	0.52	V13-84*01	30	24	59	49	46	44
V6-29*01	0.01	0.46	0.56	0.55	0.62	0.61	0.34	0.60	0.47	V6-29*01	93	44	36	36	40	35
V15-103*01	0.25	0.24	0.44	0.47	0.92	0.46	0.31	0.62	0.46	V15-103*01	56	59	41	42	26	43
V4-63*01	1.04	0.78	0.13	0.19	0.39	0.17	0.65	0.25	0.45	V4-63*01	23	26	62	64	49	63
V4-86*01	0.29	0.44	0.33	0.39	0.50	0.41	0.35	0.43	0.39	V4-86*01	53	46	51	50	44	45
V4-70*01	0.43	0.37	0.39	0.32	0.34	0.49	0.40	0.38	0.39	V4-70*01	46	53	45	53	53	40
V8-28*01	0.54	0.28	0.39	0.36	0.35	0.36	0.40	0.36	0.38	V8-28*01	36	57	46	51	52	48
V8-18*01	1.79	0.04	0.05	0.07	0.04	0.05	0.63	0.06	0.34	V8-18*01	13	84	77	75	83	80
V14-100*01	0.29	0.29	0.31	0.30	0.43	0.26	0.30	0.33	0.31	V14-100*01	54	54	53	55	47	56
V12-41*01	0.19	0.38	0.21	0.41	0.42	0.19	0.26	0.34	0.30	V12-41*01	62	52	58	48	48	60
V12-98*01	0.22	0.10	0.38	0.74	0.08	0.14	0.23	0.32	0.28	V12-98*01	61	74	48	30	78	68
V9-123*01	0.41	0.02	0.02	0.09	0.03	1.07	0.15	0.40	0.28	V9-123*01	48	92	87	73	86	25
V4-80*01	0.12	0.23	0.30	0.50	0.28	0.21	0.22	0.33	0.28	V4-80*01	69	60	54	40	56	59
V12-89*01	0.18	0.21	0.27	0.57	0.23	0.16	0.22	0.32	0.27	V12-89*01	63	61	56	35	59	66
V5-45*01	0.16	0.42	0.11	0.22	0.38	0.30	0.23	0.30	0.27	V5-45*01	64	47	69	61	50	52
V8-21*01	0.16	0.29	0.33	0.23	0.16	0.36	0.26	0.25	0.26	V8-21*01	64	56	52	60	66	47
V1-99*01	0.36	0.47	0.11	0.16	0.11	0.19	0.31	0.16	0.23	V1-99*01	50	42	66	66	76	61
V6-14*01	0.41	0.14	0.23	0.13	0.31	0.16	0.26	0.20	0.23	V6-14*01	49	69	57	69	55	66

B

	G1	G2	G3	F1	F2	F3
V4-74*01	0.49	0.25	0.12	0.21	0.14	0.10
V8-16*01	0.09	0.39	0.18	0.13	0.17	0.31
V2-112*01	0.14	0.11	0.51	0.22	0.18	0.12
V4-79*01	0.55	0.15	0.07	0.12	0.10	0.22
V1-88*01	0.09	0.06	0.39	0.12	0.13	0.28
V18-36*01	0.00	0.15	0.20	0.17	0.22	0.17
V4-50*01	0.08	0.29	0.10	0.09	0.20	0.13
V3-1*01	0	0.06	0.12	0.29	0.21	0.18
V1-122*01	0.22	0.19	0.10	0.06	0.17	0.05
V1-133*01	0.05	0.11	0.09	0.07	0.35	0.11
V4-69*01	0.09	0.21	0.03	0.06	0.13	0.22
V9-129*01	0.03	0.14	0.02	0.42	0.07	0.03
V7-33*01	0.09	0.09	0.10	0.06	0.20	0.12
V4-58*01	0.12	0.11	0.11	0.04	0.14	0.13
V14-130*01	0.04	0.21	0.07	0.15	0.13	0.04
V11-125*01	0.15	0.09	0.12	0.07	0.12	0.07
V4-54*01	0.01	0.01	0.01	0.45	0.01	0.04
V4-90*01	0.02	0.04	0.06	0.03	0.15	0.17
V12-38*01	0.03	0.19	0.05	0.04	0.06	0.07
V10-95*01	0.06	0.09	0.03	0.12	0.07	0.04
V3-9*01	0.23	0.02	0.05	0.01	0.04	0.01
V4-71*01	0.01	0.05	0.00	0.00	0.24	0.01
V4-62*01	0.05	0.08	0.01	0.01	0.01	0.06
V8-23-1*01	0.02	0.10	0.00	0.02	0.03	0.05
V4-51*01	0.03	0.03	0.04	0.01	0.06	0.02
V5-37*01	0.02	0.05	0.03	0.02	0.04	0.02
V4-81*01	0.02	0.03	0.03	0.02	0.03	0.04
V3-3*01	0.01	0.03	0.04	0.02	0.03	0.02
V4-78*01	0.02	0.03	0.01	0.02	0.02	0.04
V4-92*01	0.01	0.03	0.02	0.03	0.02	0.03
V8-34*01	0.01	0.01	0.02	0.02	0.03	0.02
V1-35*01	0.00	0.03	0.04	0.01	0.01	0.01
V4-73*01	0.01	0.01	0.01	0.01	0.03	0.02
V1-132*01	0.01	0.01	0.00	0.01	0.01	0.01
V1-131*01	0.00	0.02	0.01	0.01	0.00	0.00
V8-26*01	0.02	0.00	0	0.00	0.00	0.00
V20-101-2*01	0.00	0.00	0.01	0.01	0.00	0.00

	Avg G	Avg F	Avg T
V4-74*01	0.28	0.15	0.22
V8-16*01	0.22	0.20	0.21
V2-112*01	0.26	0.17	0.21
V4-79*01	0.26	0.15	0.20
V1-88*01	0.18	0.18	0.18
V18-36*01	0.12	0.18	0.15
V4-50*01	0.15	0.14	0.15
V3-1*01	0.06	0.23	0.14
V1-122*01	0.17	0.09	0.13
V1-133*01	0.08	0.18	0.13
V4-69*01	0.11	0.14	0.12
V9-129*01	0.06	0.17	0.12
V7-33*01	0.10	0.13	0.11
V4-58*01	0.11	0.10	0.11
V14-130*01	0.11	0.11	0.11
V11-125*01	0.12	0.09	0.10
V4-54*01	0.01	0.17	0.09
V4-90*01	0.04	0.12	0.08
V12-38*01	0.09	0.05	0.07
V10-95*01	0.06	0.08	0.07
V3-9*01	0.10	0.02	0.06
V4-71*01	0.02	0.08	0.05
V4-62*01	0.05	0.03	0.04
V8-23-1*01	0.04	0.03	0.04
V4-51*01	0.03	0.03	0.03
V5-37*01	0.04	0.03	0.03
V4-81*01	0.03	0.03	0.03
V3-3*01	0.03	0.02	0.03
V4-78*01	0.02	0.03	0.02
V4-92*01	0.02	0.02	0.02
V8-34*01	0.01	0.02	0.02
V1-35*01	0.03	0.01	0.02
V4-73*01	0.01	0.02	0.02
V1-132*01	0.01	0.01	0.01
V1-131*01	0.01	0.00	0.01
V8-26*01	0.01	0.00	0.01
V20-101-2*01	0.00	0.00	0.00

	G1	G2	G3	F1	F2	F3
V4-74*01	41	58	65	63	69	74
V8-16*01	71	50	61	68	64	51
V2-112*01	67	73	39	62	67	72
V4-79*01	35	68	75	70	77	57
V1-88*01	72	81	47	71	71	55
V18-36*01	98	67	60	65	60	65
V4-50*01	74	55	72	73	62	70
V3-1*01	100	80	64	56	61	62
V1-122*01	60	66	71	79	65	79
V1-133*01	76	71	73	77	51	73
V4-69*01	73	62	66	80	73	58
V9-129*01	79	70	88	46	79	87
V7-33*01	70	76	70	78	63	71
V4-58*01	68	72	67	81	70	69
V14-130*01	78	63	74	67	72	82
V11-125*01	66	77	63	76	75	76
V4-54*01	90	96	91	44	95	85
V4-90*01	84	85	76	83	68	64
V12-38*01	81	64	79	82	82	77
V10-95*01	75	78	84	72	79	83
V3-9*01	59	93	78	95	85	96
V4-71*01	88	83	98	99	58	94
V4-62*01	77	79	92	93	94	78
V8-23-1*01	85	75	99	89	89	81
V4-51*01	79	89	81	91	81	91
V5-37*01	85	82	83	90	84	90
V4-81*01	87	88	85	88	86	84
V3-3*01	91	87	82	87	88	92
V4-78*01	83	90	94	86	92	85
V4-92*01	92	91	89	84	93	88
V8-34*01	89	95	89	85	89	92
V1-35*01	99	86	80	93	96	95
V4-73*01	95	98	93	92	91	89
V1-132*01	94	96	97	96	97	97
V1-131*01	96	94	96	97	100	100
V8-26*01	82	99	100	100	98	98
V20-101-2*01	97	100	95	97	99	98

## Appendix A.7 D- and J-gene segment and constant region heat maps

A

	G1	G2	G3	F1	F2	F3	Avg G	Avg F	Avg T		G1	G2	G3	F1	F2	F3
D1-1*01	39.20	46.80	30.60	40.24	41.34	40.16	38.87	40.58	39.73	D1-1*01	1	1	2	1	1	1
Undeter	25.84	24.02	36.60	25.36	31.09	24.74	28.82	27.06	27.94	Undeter	2	2	1	2	2	2
D2-4*01	8.61	5.58	7.43	10.19	4.51	6.80	7.21	7.17	7.19	D2-4*01	3	5	4	3	5	5
D2-3*01	5.10	6.63	11.39	6.71	4.44	8.53	7.71	6.56	7.13	D2-3*01	6	4	3	5	6	3
D4-1*01	7.83	5.31	5.26	7.92	7.83	7.25	6.13	7.66	6.90	D4-1*01	4	6	5	4	3	4
D2-5*01	6.82	7.02	3.56	4.85	4.64	6.22	5.80	5.24	5.52	D2-5*01	5	3	6	6	4	6
D3-1*01	3.64	1.75	2.49	1.87	2.63	2.83	2.63	2.44	2.54	D3-1*01	7	8	7	8	7	7
D3-2*02	2.30	1.77	1.87	2.08	2.58	1.77	1.98	2.14	2.06	D3-2*02	8	7	8	7	8	8
D6-2*02	0.46	1.00	0.26	0.23	0.46	0.63	0.58	0.44	0.51	D6-2*02	9	9	10	10	9	10
D5-1*01	0.11	0	0.39	0.45	0.33	0.35	0.17	0.38	0.27	D5-1*01	10	11	9	9	10	11
D5-5*01	0.09	0.11	0.13	0.11	0.13	0.73	0.11	0.32	0.22	D5-5*01	11	10	11	11	11	9

B

	G1	G2	G3	F1	F2	F3	Avg G	Avg F	Avg T		G1	G2	G3	F1	F2	F3
J2*01	29.35	27.01	28.75	35.26	27.19	40.43	28.37	34.29	31.33	J2*01	2	2	2	1	2	1
J1*03	30.25	40.25	19.77	19.68	20.69	24.88	30.09	21.75	25.92	J1*03	1	1	4	3	3	2
J4*01	19.66	13.09	22.60	18.92	34.70	18.78	18.45	24.13	21.29	J4*01	4	4	3	4	1	3
J3*01	20.69	19.49	28.80	26.03	16.58	15.90	22.99	19.50	21.25	J3*01	3	3	1	2	4	4
U	0.04	0.16	0.08	0.11	0.84	0.01	0.09	0.32	0.21	<6	5	5	5	5	5	5

C

	G1	G2	G3	F1	F2	F3	Avg G	Avg F	Avg T		G1	G2	G3	F1	F2	F3
J5*01	37.27	48.07	32.56	39.64	31.87	40.19	39.30	37.23	38.27	J5*01	1	1	1	1	2	1
J2*01	25.08	20.37	20.98	23.99	33.69	27.43	22.15	28.37	25.26	J2*01	3	3	3	2	1	2
J1*01	26.74	21.43	28.76	21.02	21.59	20.09	25.64	20.90	23.27	J1*01	2	2	2	3	3	3
J4*01	9.77	8.45	15.71	13.49	10.65	10.54	11.31	11.56	11.44	J4*01	4	4	4	4	4	4
U	1.14	1.68	1.99	1.86	2.20	1.76	1.60	1.94	1.77	U	5	5	5	5	5	5

D

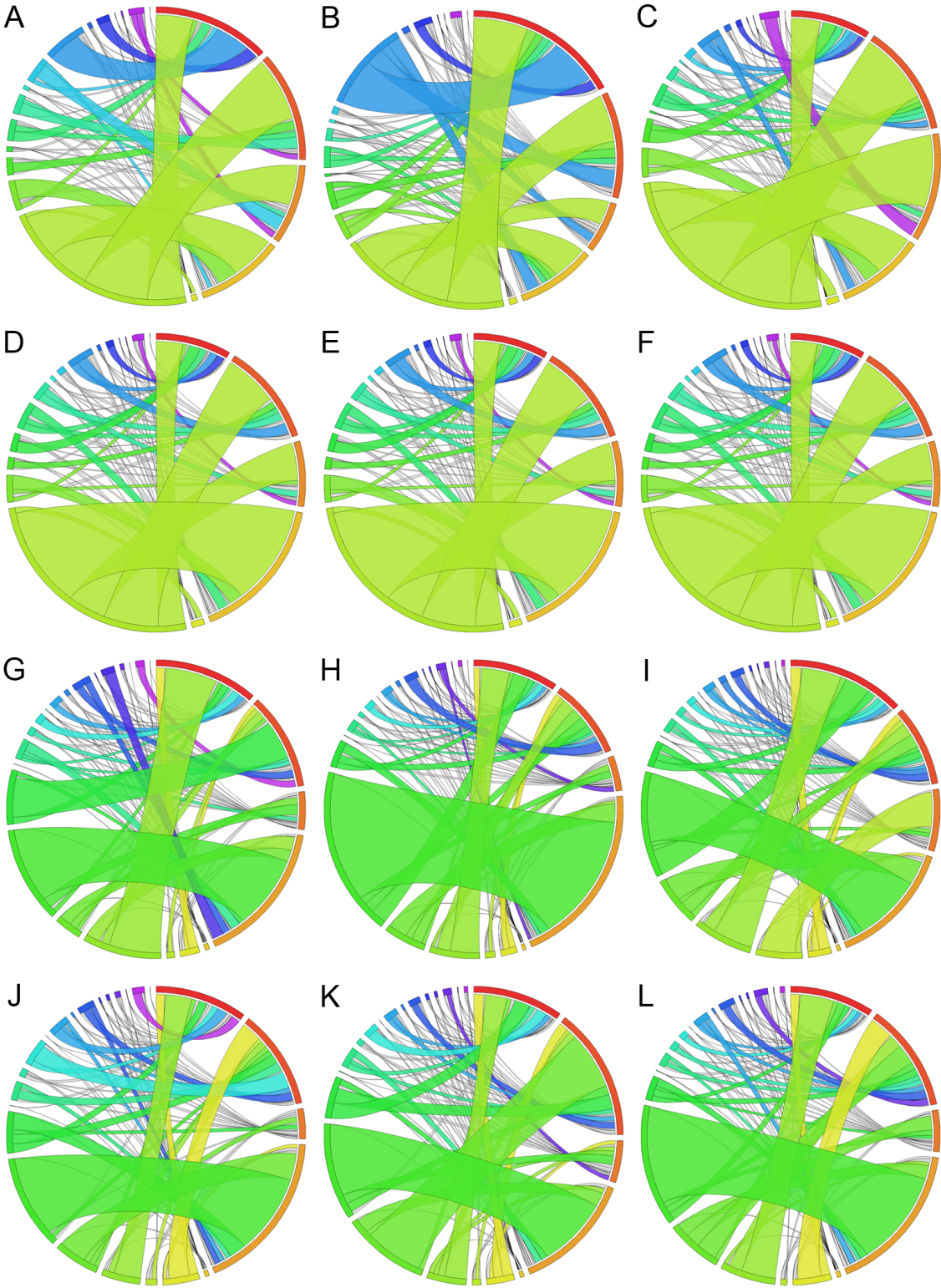
	G1	G2	G3	F1	F2	F3	Avg G	Avg F	Avg T		G1	G2	G3	F1	F2	F3
IgM	82.81	80.03	67.44	68.07	62.38	68.18	76.76	66.21	71.48	IgM	1	1	1	1	1	1
IgG	12.27	17.66	29.61	28.82	25.56	29.08	19.85	27.82	23.83	IgG	2	2	2	2	2	2
IgA	4.58	1.35	2.32	2.54	11.26	2.08	2.75	5.29	4.02	IgA	3	3	3	3	3	3
IgD	0.31	0.95	0.56	0.53	0.79	0.64	0.61	0.66	0.63	IgD	4	4	4	4	4	4
IgE	0.04	0.02	0.07	0.03	0.01	0.02	0.04	0.02	0.03	IgE	5	5	5	5	5	5

(A) D-gene segment, (B) JH-gene segment, (C) J $\kappa$ -gene segment, (D) IgH constant region usage in animals within ground (G) and flight (F) treatment groups are presented as percent of repertoire and rank. Dark red indicates higher percent of repertoire or rank moving to white, lower percent of repertoire or rank.

## **Appendix A.8 V/J combinations of individual animals**

(See following page.)

(A-C) IgH V/J pairings from G1, G2, and G3 respectively. (D-F) IgH V/J pairings from F1, F2, and F3 respectively. Circos plots are read clockwise starting at the 12 o'clock position starting with J1 (red), J2, J3, J4, U, V1 (lime green), V2, V3, V4, V5, V6, V7, V8 (light blue), V9, V10, V11, V12, V13, V14, and V15 (sliver, no color). (G-I) Igκ V/J pairings from G1, G2, and G3 respectively. (J-L) Igκ V/J pairings from F1, F2, and F3 respectively. Circos plots are read clockwise starting at the 12 o'clock position with J1 (red), J2, J4, J5, U, V1 (yellow), V2, V3, V4, V5, V6, V7 (sliver, no color), V8, V9, V10, V11, V12, V13, V14, V15, V16, V17, V18, V19 (light purple).



## **Appendix A.9 Top V(D)J gene family combinations**

(See following page.)

(A, B) Top 5 V/D/J combinations for IgH ground (A) and flight (B) mice with average and standard deviation (SD) per treatment group. (C, D) Top 5 V/J combinations for IgH ground (C) and flight (D) mice. (E, F) Top 5 V/J combinations for Igκ ground (E) and flight (F) mice with average and standard deviation (SD) per treatment group. Gene combinations are color coded to show overlap within the top five most common combinations per treatment group. Combinations found within all three mice are blue, within two mice are green, and unique to a single mouse as white.

A

G1	%	G2	%	G3	%
V9/D1/J1	8.40	V9/D1/J1	10.5	V1/U/J3	13.1
V1/U/J4	7.90	V1/D1/J1	9.89	V1/U/J2	5.92
V1/D1/J2	6.95	V1/D1/J2	4.83	V1/D2/J2	5.61
V1/D1/J1	5.50	V1/U/J2	3.81	V1/D1/J2	4.83
V1/U/J2	3.93	V1/U/J4	3.25	V1/U/J4	4.23

Avg G	%	SD
V9/D1/J1	6.86	4.63
V1/D1/J1	6.15	3.46
V1/U/J3	5.97	6.17
V1/D1/J2	5.54	1.22
V1/U/J4	5.13	2.45

	Shared among 3 mice
	Shared among 2 mice
	Unique to 1 mouse

B

F1	%	F2	%	F3	%
V1/D1/J2	9.63	V1/D1/J4	20.3	V1/D1/J2	9.09
V3/D1/J3	8.02	V1/U/J3	4.87	V9/D1/J1	6.10
V1/D1/J1	6.82	V1/U/J2	4.59	V1/U/J2	5.45
V1/U/J2	6.09	V1/D1/J2	4.30	V1/D2/J2	4.90
V5/D2/J2	4.95	V1/D1/J1	3.69	V1/D1/J3	3.59

Avg F	%	SD
V1/D1/J4	8.27	10.44
V1/D1/J2	7.68	2.93
V1/U/J2	5.38	0.75
V1/D1/J1	4.68	1.86
V1/U/J3	3.66	1.06

C

G1	%	G2	%	G3	%
V1/J2	16.80	V1/J1	14.3	V1/J3	19.1
V1/J4	11.70	V1/J2	12.3	V1/J2	18.9
V1/J3	10.6	V9/J1	12.3	V1/J4	11
V1/J1	9.74	V1/J4	9.27	V1/J1	7.65
V9/J1	8.75	V1/J3	6.22	V2/J4	4.19

D

F1	%	F2	%	F3	%
V1/J2	20.7	V1/J4	26.3	V1/J2	22.7
V1/J1	12.2	V1/J2	14	V1/J1	10.20
V1/J4	12	V1/J3	9.82	V1/J3	9.99
V1/J3	10.3	V1/J1	6.99	V1/J4	9.11
V3/J3	10.3	V2/J4	3.58	V9/J1	6.3

E

G1	%	G2	%	G3	%
V5/J5	20.0	V5/J5	34.6	V5/J5	18.2
V3/J1	13.9	V3/J1	7.70	V3/J1	10.6
V6/J2	10.5	V4/J5	4.97	V2/J4	9.32
V3/J5	3.67	V5/J1	4.31	V5/J1	6.66
V4/J5	3.15	V3/J2	3.60	V4/J5	6.31

Avg G	%	SD
V5/J5	24.3	9.02
V3/J1	10.7	3.09
V6/J2	4.86	4.90
V4/J5	4.81	1.59
V5/J1	4.05	2.74

F

F1	%	F2	%	F3	%
V5/J5	21.7	V5/J5	19.0	V5/J5	25.8
V4/J5	7.28	V4/J2	15.5	V3/J1	9.07
V3/J1	7.06	V3/J1	5.81	V1/J2	6.26
V1/J2	5.49	V6/J1	4.22	V4/J2	5.41
V6/J5	4.55	V6/J2	3.93	V4/J4	3.8

Avg F	%	SD
V5/J5	22.2	3.40
V4/J2	7.43	7.30
V3/J1	7.31	1.64
V4/J5	4.88	2.09
V1/J2	4.73	2.02



# Appendix B - Specific Procedures for Repertoire Analysis

## Appendix B.1 Immunoglobulin reference sequences

**Note:** The following is an outline of how reference sequences were obtained. It is not necessary to repeat these steps as final references are stored in CLC for future mappings.

- Reference sequences were obtained from [www.imgt.org/vquest/refseqh.html](http://www.imgt.org/vquest/refseqh.html) (current as of 06/2017).
- Under “IMGT/V-QUEST reference directory sets”, the mouse IGκV fasta list was selected within the “F + ORF + all P” column of the “IG ‘V-Region’, ‘D-Region’, ‘J-region’, C-Gene’ sets” table.

IG "V-REGION", "D-REGION", "J-REGION", "C-GENE exon" sets

Groups	F+ORF+all P	F+ORF+in-frame P		F+ORF+in-frame P with IMGT gaps	
	Nucleotides	Nucleotides	Amino acids	Nucleotides	Amino acids
IGHV	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque
IGHD	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque		
IGHJ	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rabbit Rainbow trout, Rhesus monkey, Pig, Zebrafish, Platypus, Alpaca, Crab-eating macaque		
IGHC	Human Mouse Rat, Rainbow trout, Pig, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rainbow trout, Pig, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rainbow trout, Pig, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rainbow trout, Pig, Platypus, Alpaca, Crab-eating macaque	Human Mouse Rat, Rainbow trout, Pig, Platypus, Alpaca, Crab-eating macaque
IGKV	Human Mouse, Pig, Rat, Rabbit, Rhesus monkey	Human Mouse, Pig, Rat, Rabbit, Rhesus monkey	Human Mouse, Pig, Rat, Rabbit, Rhesus monkey	Human Mouse, Pig, Rat, Rabbit, Rhesus monkey	Human Mouse, Pig, Rat, Rabbit, Rhesus monkey

**Nucleotide sequences for F+ORF+all P alleles including orphans**

---

The FASTA header contains 15 fields separated by '|':

- IMGT/LIGM-DB accession number(s)
- gene and allele name
- species
- functionality
- exon(s), region name(s), or extracted label(s)
- start and end positions in the IMGT/LIGM-DB accession number(s)
- number of nucleotides in the IMGT/LIGM-DB accession number(s)
- codon start, or 'NR' (not relevant) for non coding labels
- +n: number of nucleotides (nt) added in 5' compared to the corresponding label extracted from IMGT/LIGM-DB
- +n or -n: number of nucleotides (nt) added or removed in 3' compared to the corresponding label extracted from IMGT/LIGM-DB
- +n, -n, and/or nS: number of added, deleted, and/or substituted nucleotides to correct sequencing errors, or 'not corrected' if non corrected sequencir
- number of amino acids (AA): this field indicates that the sequence is in amino acids
- number of characters in the sequence: nt (or AA)+IMGT gaps=total
- partial (if it is)
- reverse complementary (if it is)

---

Number of results = 210

```
>A1231284|IGKV1-108*01|Mus musculus_C3H|P|V-REGION|980..1278|299 nt|1| || |299-0=299| | |
tattaaaaaatcagaagattgctcgaagttgtacctgattctgtactattcttca
ttagcacatctagtaagcctctgtcacacgaattggaattcttattgggtggcacc
ttgcagaagccaggcagctcttacaactctgattcattgagtttccaaacgaattct
ggggttccagacagttcagtgccagtgattcagggacagatttcacacttaagtcagc
```

1. Individual FASTA sequences were pasted into Notepad and all line breaks within the nucleotide sequence were removed.
2. Files were named after the gene segment and allele (For example, IGκV1-52-01) and saved in the file format “.fasta”.
3. Fasta files are stored on the lab computer under the following location:  
C:\Users\ChapesLab\Documents\Claire\IGKV FASTA
4. Because sequences obtained from IMGT are representative of multiple mouse strains, a C57BL/6 specific IGκV list was created by selecting only IMGT obtained gene segments that were also found in the within *Mus musculus* (GRCm38.p4, C57BL/6J assembly) Igκ locus within the NCBI gene database: <https://www.ncbi.nlm.nih.gov/gene/243469> (current as of 6/2017) .
5. In some instances, multiple alleles for a specific C57BL/6 gene segment were available from IMGT. In these cases, the proper allelic designation was identified by reviewing the number of assignments to the possible alleles for a gene segment by the IMGT HighV-Quest tool. An excerpt from the decision making take is shown below:

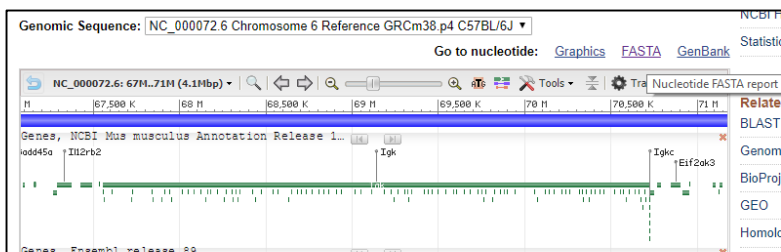
<b>Allelic Designation for IGKV</b>					
<b>NCBI Designation</b>	<b>IMGT Alleles</b>	<b>Cell Count (Frequency) N=5394</b>	<b>Tissue Count (Frequency) N=9192</b>	<b>SS Count (Frequency) N=18462</b>	<b>Decision</b>
<b>IGKV2-137</b>	*01	97 (.017)	196 (.021)	382 (.021)	Use *01
	*02	0	0	0	
<b>IGKV14-126</b>	*01	49 (.009)	141 (.015)	252 (.014)	Use *01
	*02	0	0	0	
<b>IGKV1-117</b>	*01	446 (.083)	748 (.081)	1668 (.090)	Use *01
	*02	1 (.000)	12 (.001)	5 (.000)	

6. Predominant alleles were assumed to be the C57BL/6J allele retained in the C57BL/6J reference list.
7. Tied alleles were also retained.

8. The C57BL/6J reference list was further refined to exclude pseudogenes, which were identified by the letter “P” within the header of fasta files obtained from IMGT.
9. Fasta read files were imported into CLC Genomics Workbench by selecting “Import”, “FASTA Read Files” and selecting all final C57BL/6J gene segment references within C:\Users\ChapesLab\Documents\Claire\IGKV FASTA.
10. The final C57BL6/J IGKV gene segment references are located in CLC Genomics Workbench under the file path: CLC\_Data>General>FASTA>IgK Fastas>V Kappa Chain B6 only.

**Note: Mapping to the entire Igk locus allows for the capture sequencing reads that are not otherwise captured by individual IGKV reference sequences. The following steps outline how the Igk locus reference was obtained.**

11. The *Mus musculus* (GRCm38.p4, C57BL/6J assembly) Igκ locus was obtained from the NCBI gene database <https://www.ncbi.nlm.nih.gov/gene/243469> (current as of 6/2017)
12. A FASTA file of the located was generated by selecting the FASTA link on the upper right-hand side of the genome viewer.



13. The sequence was copied into a Notepad file and line breaks within the nucleotide sequence were removed. The file was named “IGK locus” and saved as a fasta file within C:\Users\ChapesLab\Documents\Claire\IGKV FASTA.

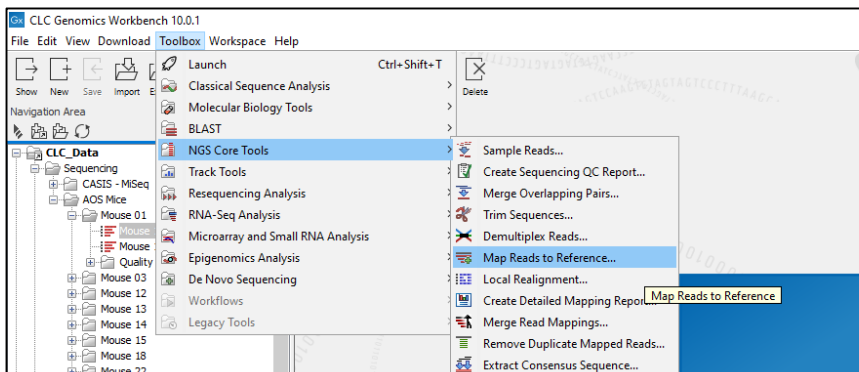
14. The IGκ locus was imported into CLC Genomics workbench and is located under the file path: CLC\_Data>General>FASTA>IgK FASTAS.

## Appendix B.2 Mapping

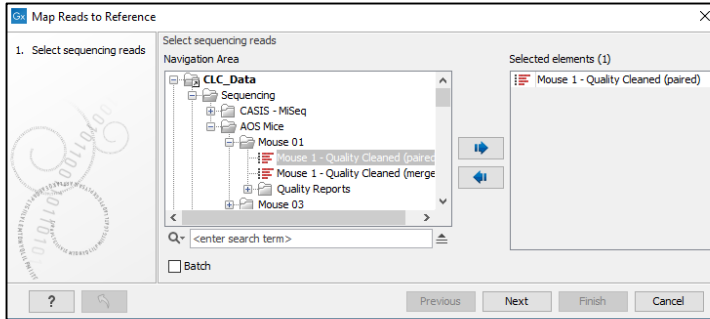
1. Within CLC Genomics Workbench, select the “Sequencing” folder and the folder of the project and animal to be mapped (for example: Sequencing>AOS Mice>Mouse 01).

**Note:** Within the animal folder there are two quality cleaned sequence lists: paired and merged. Both paired and merged files should be independently mapped to 1) C57BL/6J IGκV references and 2) the IGκ locus. The following steps outline how to map one paired sequences to IGκV references. These steps should be repeated to also map paired sequences to the IGκ locus, merged sequences to IκV references, and merged sequences to the IGK locus.

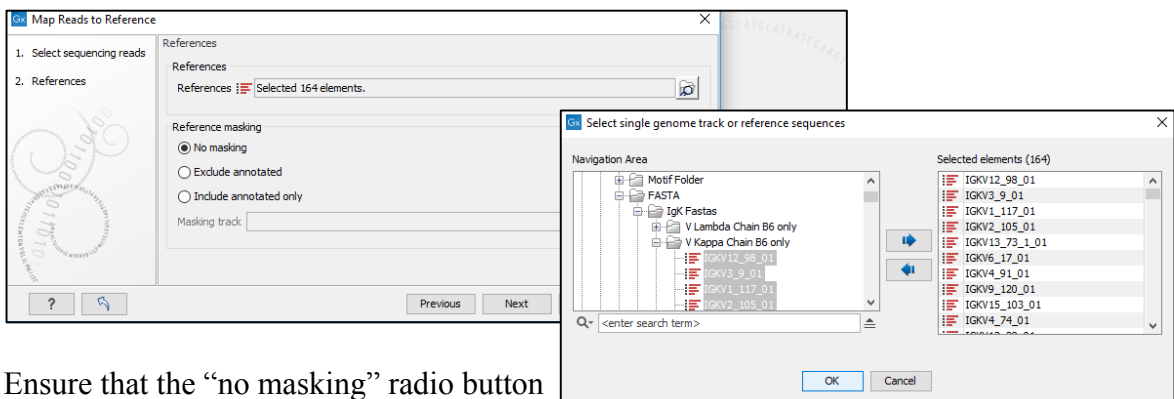
2. Under “Toolbox”, select “NGS Core Tools”, and “Map Reads to Reference”.



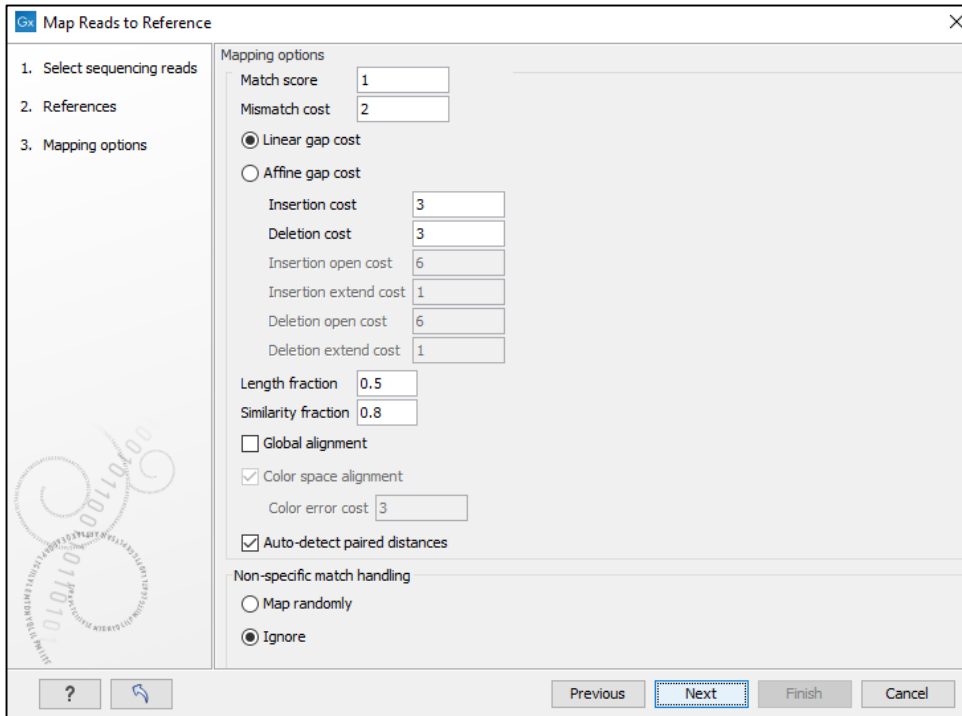
3. Within the dialogue box, select the desired sequence list for mapping. If the list is highlighted within the navigation area prior to selecting the “Map Reads to Reference” tool, the desired sequencing list will already appear selected within the dialogue box. Once selected, select “Next”.



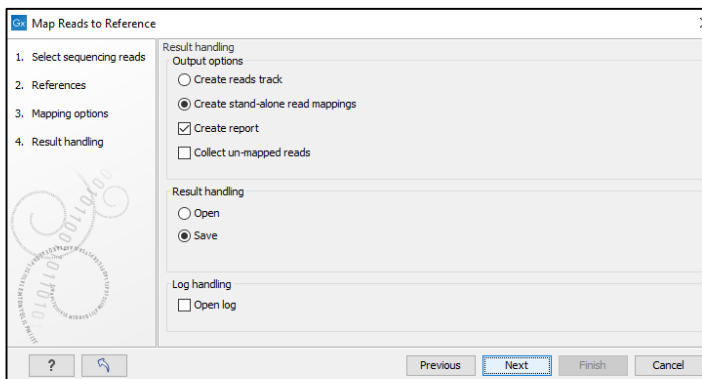
4. Indicate references for mapping. To change the references, select the browse and select element button to the right of the reference bar.



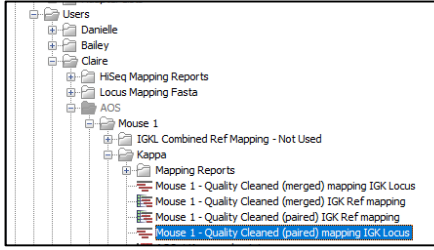
5. Ensure that the “no masking” radio button is checked under the “reference masking” section and select “Next”.
6. Ensure that the mapping options match what is selected in the following screenshot and select “Next”.



7. Ensure that the output options match what is selected in the following screenshot and select “Next”.

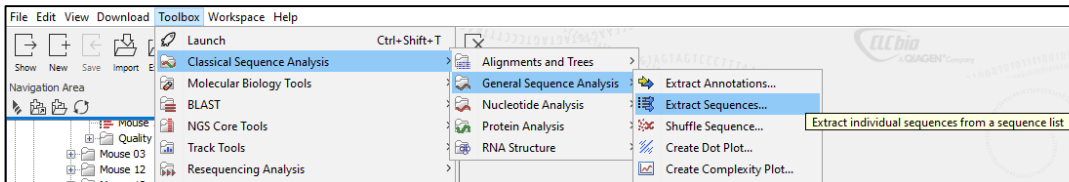


8. On the final screen, select an appropriate location to save the output files and select “Finish”. It is recommended to save all mapping output files for an animal within the same folder as shown below:

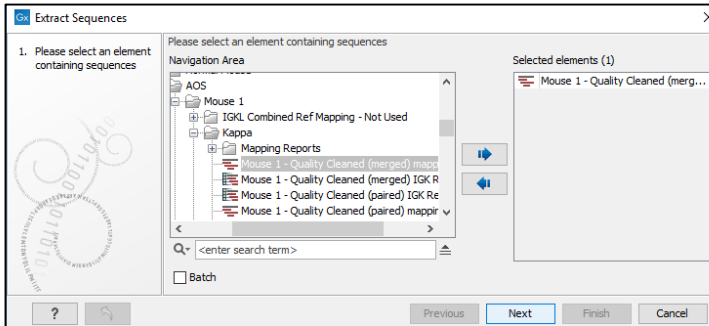


9. Once mapping to IGκV references and the IGκ locus is complete for both paired and merged reference lists, mapped sequences are extracted and combined for submission to the IMGT HighV-Quest tool.

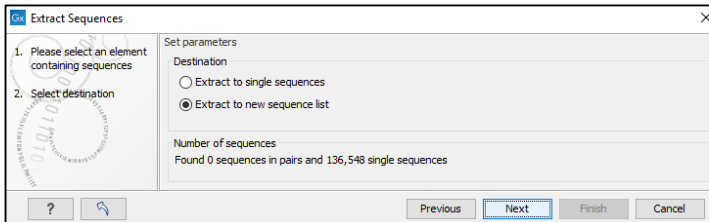
10. Extract sequences individually from all mapped sequence lists by selecting “Toolbox”, “Classical Sequence Analysis”, “General Sequence Analysis” and “Extract Sequences”.



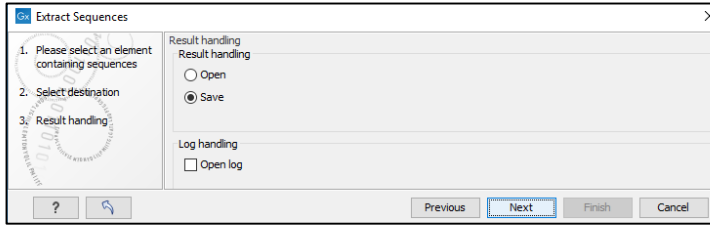
11. Ensure the desired mapped sequence list is selected and select “Next”.



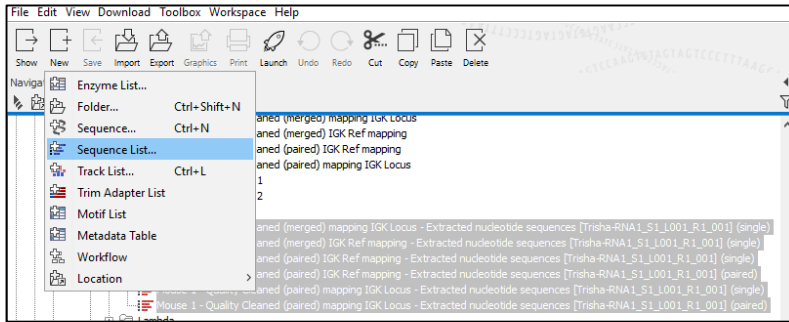
12. Ensure that “Extract to new sequence list” is selected and select “Next”.



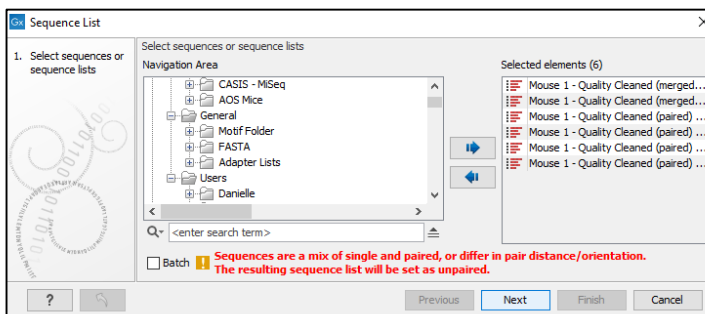
13. Ensure that “Save” is selected under “Result handling” and select “Next”.



14. Save in the same folder that the mapped sequence is located and select “Finish”.
15. Repeat sequence extractions for all mapped sequences. Two outputs will be created for paired mapped sequences (single and paired), totaling to six extraction outputs overall (2 merged, 4 paired).
16. Select all six sequence extraction outputs and select “New”, “Sequence List” to combine.



17. A warning will appear that single and paired sequences are being combined. This is okay. Select “Next” to continue.



18. Select “Save” under “Result handling” and select “Next”.

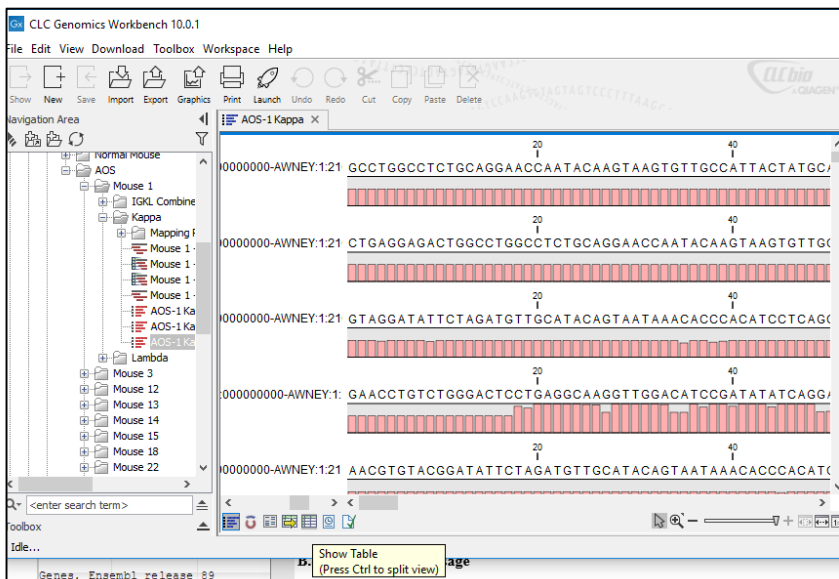




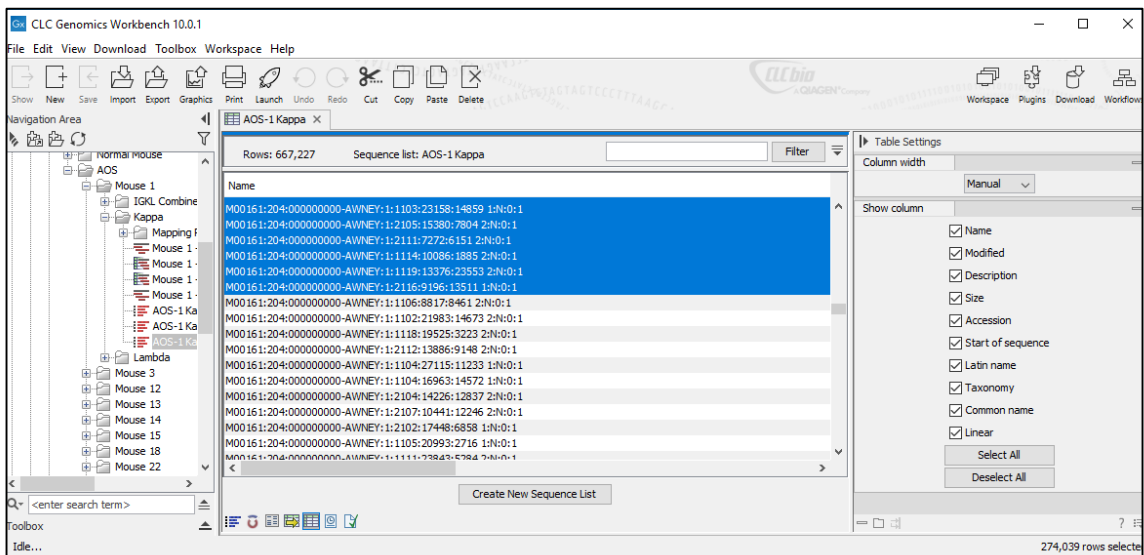
19. Save in the same folder that mapped sequences are located and select “Finish”.
20. Rename the new sequence list to reflect the experiment and animal identification number followed by the immunoglobulin chain that sequences were mapped to (For example, “AOS-1 Kappa”).

**Note:** Because IMGT can only accommodate submissions containing less than 500,000 sequencing reads, it is possible that the combined sequencing lists must be broken up and submitted into multiple parts.

21. To determine whether the sequence list must be divided into parts, double click the sequencing list and then select the “Show Table” icon located at the bottom of the sequence viewer.



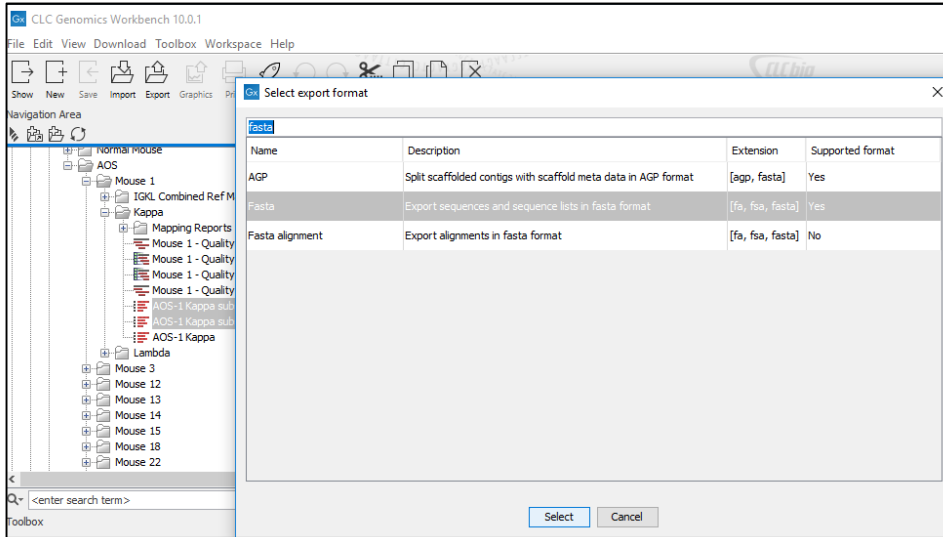
22. In the table view, the number of rows displayed indicates the number of sequencing reads within the list. In the screenshot below the list contains 667,227 sequences. Because over 500,000 sequences are present, the list must be broken into subsets by highlighting rows and selecting “Create New Sequence List”. The number of sequences for each subset must be estimated based on the number of reads. In this case it is sufficient to only create two subsets. Select the first row and hold the “Shift” key while scrolling through the table. While still holding the “shift” key, select a second row to highlight all intervening rows.



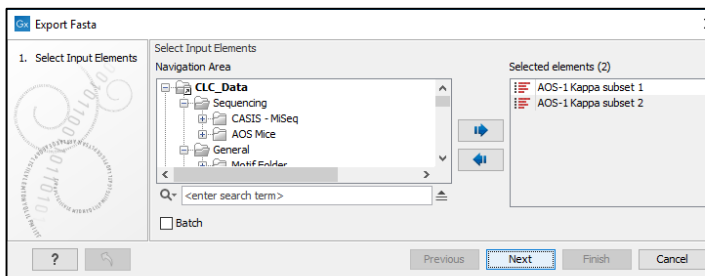
23. The resultant subset should be saved with a subset number. Select the row immediately under the last row highlighted from the first subset and highlight the rest of the rows to be included in the next subset.

24. Select “Create New Sequence List” and save the resultant subset with the subset number.

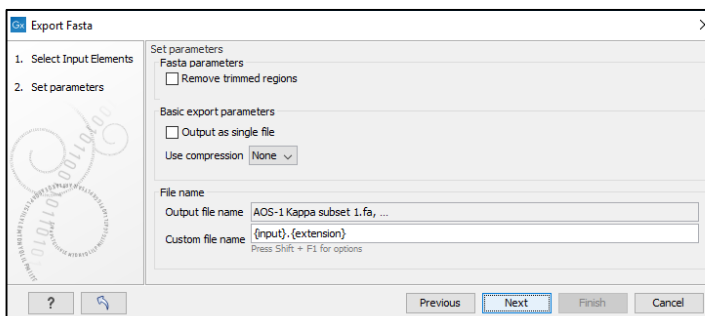
25. Export the combined sequence list or sequence list subsets by highlighting the list(s) and selecting “Export”, and selecting the FASTA export format.



26. Confirm that the appropriate lists have been selected for export and select “Next”.



27. Ensure that “remove timed regions” and “Output as single file” boxes are left unchecked and select “Next”.



28. Save fasta files in a folder on the computer for subsequent upload to the IMGT HighV-Quest tool.

## Appendix B.3 IMGT submission

1. Log in to the IMGT HighV-Quest tool: <http://imgt.org/HighV-QUEST/login.action> (Current as of 06/2017).
2. Under the “IMGT/HighV-Quest Search page”, enter the analysis details by specifying an “Analysis title”, specify *Mus musculus* under “Species”, IGκ under “Receptor type or locus”, and indicate whether sequences are from an individual.
3. Upload sequences in fasta format by selecting the desired list from the CLC export.
4. Under the section “A. Detailed View” select the “check all” option.
5. Select the “Start” button to upload the file and cue the analysis.

The screenshot shows the IMGT/HighV-QUEST web interface. At the top, there is a navigation bar with links for 'Login: ceward@ksu.edu', 'IMGT/HighV-QUEST Search page', 'Analysis history', 'Launch statistics', 'Statistics history', 'IMGT/StatClonotype NEW!', 'Help', and 'Logout'. Below the navigation bar, there is a version information section: 'IMGT/HighV-QUEST version: 1.5.5 (9 June 2017) IMGT/QUEST version: 3.4.7 (8 June 2017) IMGT/QUEST reference directory release: 201723-4 (8 June 2017)'. A yellow box highlights a 'Citing IMGT/HighV-QUEST:' section with several references. The main form area contains the following fields and options:

- Analysis title:** AOS-1 Kappa Subset 1 (50 characters or less)
- Species:** Mus musculus (house mouse)
- Receptor type or locus:** IGκ
- Sequences are from a single individual:** Yes
- Upload sequences in FASTA format:** Browse... No file selected. Submission up to 500000 sequences. For more than 150000 sequences, the individual files are not provided.
- E-mail notifications:**  when analysis is queued  when analysis is completed [Check all](#) | [None](#)
- Start** button

Below the form, there is a 'Display results' section with the following options:

- A. Detailed View** (options for submission < 150 000 sequences)
  - Include individual result files:  Yes  No
  - Nb of nucleotides per line in alignments: 60
  - Nb of aligned reference sequences: 5
- 13 checkboxes for various analysis options, all of which are checked:
  - 1. Alignment for V-GENE
  - 2. Alignment for D-GENE
  - 3. Alignment for J-GENE
  - 4. Results of IMGT/JunctionAnalysis
    - with full list of eligible D-GENES
    - without list of eligible D-GENES
  - 5. Sequence of the JUNCTION (nt and AA)
  - 6. V-REGION alignment
  - 7. V-REGION translation
  - 8. V-REGION protein display
  - 9. V-REGION mutation and AA change table
  - 10. V-REGION mutation and AA change statistics
  - 11. V-REGION mutation hotspots
  - 12. Sequences of V, V-J or V-D-J-REGION (nt and AA) with gaps in FASTA
  - 13. Annotation by IMGT/Automat

At the bottom of the form, there are links for 'Check all', 'None', and 'Default'.


6. If applicable, submit all subsets of mapped sequence lists separately.
7. When the analysis results are available, results will be cataloged on the “Analysis history” page.





































Login: ceward@ksu.edu [IMGT/HighV-QUEST Search page](#) [Analysis history](#) [Launch statistics](#) [Statistics history](#) [IMGT/StatClonotype \*\*NEW!\*\*](#) [Help](#) [Logout](#)

IMGT/HighV-QUEST version: [1.5.5](#) (9 June 2017) IMGT/V-QUEST version: [3.4.7](#) (8 June 2017) IMGT/V-QUEST reference directory release: [201723-4](#) (8 June 2017)

**Citing IMGT/HighV-QUEST:**  
 Alamyar, et al. IMGT/HighV-QUEST: The IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res. 8:1-2 (2012). LIGM:400 [PMID:22647994](#) [PDF](#)  
 Alamyar E., et al., Methods Mol. Biol. 882:569-604 (2012). [PMID:22665256](#) LIGM:404  
 Li S., et al. IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype, clonal expression evaluation diversity and next generation repertoire immunoprofiling. Nat. Commun. 4:2333 (2013). [Open access](#) [PMID:23995877](#) LIGM:419  
 Giudicelli V., et al., Autoimmun Infect Dis. 1(1) (2015). doi:10.16966/aidoa.103. [Free Article](#) LIGM:448

Here is information about your analysis history

 To launch statistical analysis on completed jobs (now with IMGT clonotype (AA)), click on 'Launch statistics' on the menu bar.

Title	User	Status	Submission date	Nb of sequences	IMGT/V-QUEST reference directory		Actions
					Species	Receptor type (or locus)	
AOS-15 Kappa Part 1 of 1	Claire Ward	completed 	2017-05-26 18:30:37 <a href="#">CET</a>	453385	Mus_musculus	IGK	(42.87 MB, 12 files)  
AOS-14 Kappa Part 2	Claire Ward	completed 	2017-05-26 18:24:08 <a href="#">CET</a>	430230	Mus_musculus	IGK	(49.28 MB, 12 files)  
AOS-14 Kappa Part 1	Claire Ward	completed 	2017-05-26 18:16:25 <a href="#">CET</a>	338718	Mus_musculus	IGK	(29.17 MB, 12 files)  
AOS-13 Kappa Subset 2	Claire Ward	completed 	2017-05-26 18:01:23 <a href="#">CET</a>	340006	Mus_musculus	IGK	(30.37 MB, 12 files)  
AOS-13 Kappa Subset 1	Claire Ward	completed 	2017-05-26 17:56:39 <a href="#">CET</a>	281255	Mus_musculus	IGK	(24.22 MB, 12 files)  
AOS-12 Kappa Subset 3	Claire Ward	completed 	2017-05-26 17:53:17 <a href="#">CET</a>	415689	Mus_musculus	IGK	(53.72 MB, 12 files)  
AOS-12 Kappa Subset 2	Claire Ward	completed 	2017-05-26 17:48:58 <a href="#">CET</a>	358850	Mus_musculus	IGK	(35.61 MB, 12 files)  
AOS-12 Kappa Subset 1	Claire Ward	completed 	2017-05-26 17:43:42 <a href="#">CET</a>	371284	Mus_musculus	IGK	(38.66 MB, 12 files)  
AOS-3 Kappa Subset 2	Claire Ward	completed 	2017-05-26 17:38:43 <a href="#">CET</a>	367632	Mus_musculus	IGK	(32.17 MB, 12 files)  
AOS-3 Kappa Subset 1	Claire Ward	completed 	2017-05-26 17:33:02 <a href="#">CET</a>	302207	Mus_musculus	IGK	(34.49 MB, 12 files)  
AOS-1 Kappa Subset 2	Claire Ward	completed 	2017-05-26 17:28:03 <a href="#">CET</a>	357920	Mus_musculus	IGK	(27.43 MB, 12 files)  
AOS-1 Kappa Subset 1	Claire Ward	completed 	2017-05-26 17:21:45 <a href="#">CET</a>	309307	Mus_musculus	IGK	(35.93 MB, 12 files)  

8. Download the zipped folder of IMGT output files and save the file.
9. If applicable, ensure that zipped IMGT output files are downloaded for all subsets.

## Appendix B.4 Cleaning IMGT output data

**Note:** The following list includes fields from IMGT output files have been determined as necessary for current and future downstream analyses. Because all analyses are to be performed on data that have undergone duplicate read removal, it is important to consolidate the fields of interest into a single excel spreadsheet prior to initial cleaning.

## **IMGT Output Files Used for Repertoire Analysis**

**Output 1: Use all EXCEPT F-I, K-M, O-R, V, X, Z, AD (delete the following)**

- (F) V-REGION identity %
- (G) V-REGION identity nt
- (H) V-REGION identity % (with ins/del events)
- (I) V-REGION identity nt (with ins/del events)
- (K) J-REGION score
- (L) J-REGION identity %
- (M) J-REGION identity nt
- (O) D-REGION reading frame
- (P) CDR1-IMGT length
- (Q) CDR2-IMGT length
- (R) CDR3-IMGT length
- (V) JUNCTION frame
- (X) Functionality comment
- (Z) J-GENE and allele comment
- (AD) Deleted n nt

**Output 2: Use columns J-R (include the following)**

- (J) FR1-IMGT
- (K) CDR1-IMGT
- (L) FR2-IMGT
- (M) CDR2-IMGT
- (N) FR3-IMGT
- (O) CDR3-IMGT
- (P) JUNCTION
- (Q) J-REGION
- (R) FR4-IMGT

**Output 3: Not used**

**Output 4: Not used**

**Output 5: Not used**

**Output 6: Use columns J-AD (include the following)**

- (J) 3'V-REGION

(K) P3'V  
(L) N-REGION  
(M) N1-REGION  
(N) P5'D  
(O) D-REGION  
(P) P3'D  
(Q) P5'D1  
(R) D1-REGION  
(S) P3'D1  
(T) N2-REGION  
(U) P5'D2  
(V) D2-REGION  
(W) P3'D2  
(X) N3-REGION  
(Y) P5'D3  
(Z) D3-REGION  
(AA) P3'D3  
(AB) N4-REGION  
(AC) P5'J  
(AD) 5'J-REGION

(F) FR1-IMGT  
(G) CDR1-IMGT  
(H) FR2-IMGT  
(I) CDR2-IMGT  
(J) FR3-IMGT  
(K) CDR3-IMGT

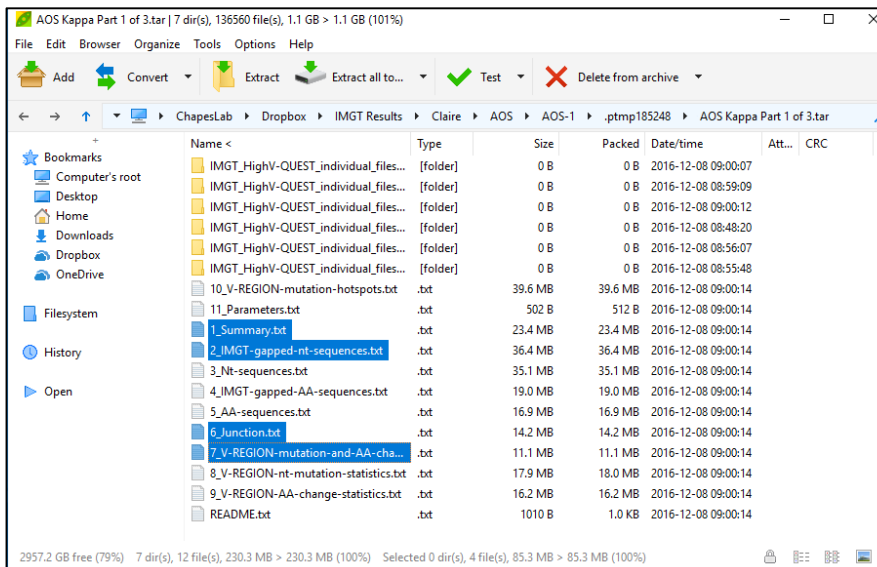
**Output 8: Not Used**

**Output 10: Not used**

**Output 11: Not used**

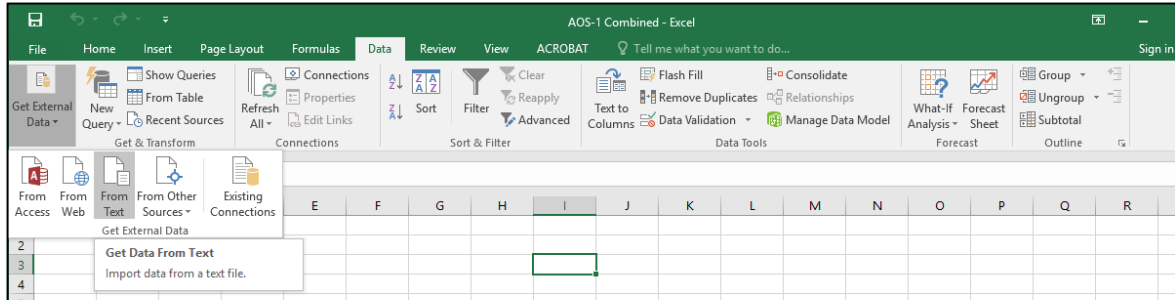
**Output 7: Use columns F-K (include the following)**

1. IMG T HighV-Quest files are stored in a compressed folder that can be opened with the free program PeaZip. In the case of AOS spleen data, these files are stored in the DropBox under the following file pathway: C:\Users\ChapesLab\Dropbox\IMG T Results\Claire\AOS
2. For convenience, double click on the IMG T output data files to unzip the folder and temporarily drag the needed files onto the desktop (outputs: 1, 2, 6, 7). For an example the files AOS-1 provided. Located at C:\Users\ChapesLab\Dropbox\IMG T Results\Claire\AOS\AOS-1

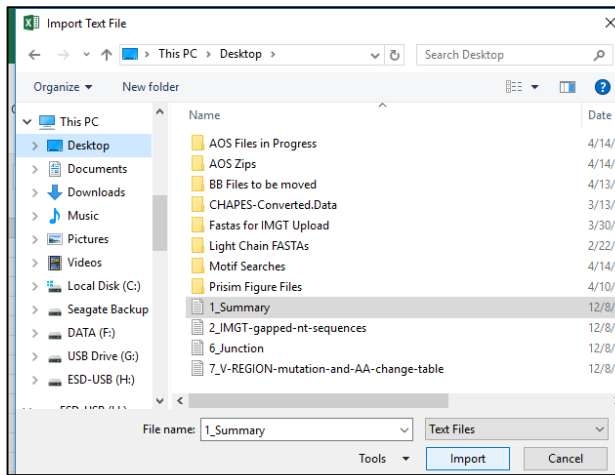


3. If the file is broken into multiple IMG T submissions (AOS-1 is broken into two separate submissions), it is important to either rename the files to include an identifier such as “AOS-1 part 1” to each of the transferred IMG T outputs or delete the transferred outputs from the desktop once imported into Microsoft Excel so that no output files are inadvertently mixed up.
4. Open Microsoft Excel spreadsheet and save the spreadsheet in the appropriate folder.
5. Within the spreadsheet, select the “Data” tab, “Get External Data” and “From Text”.

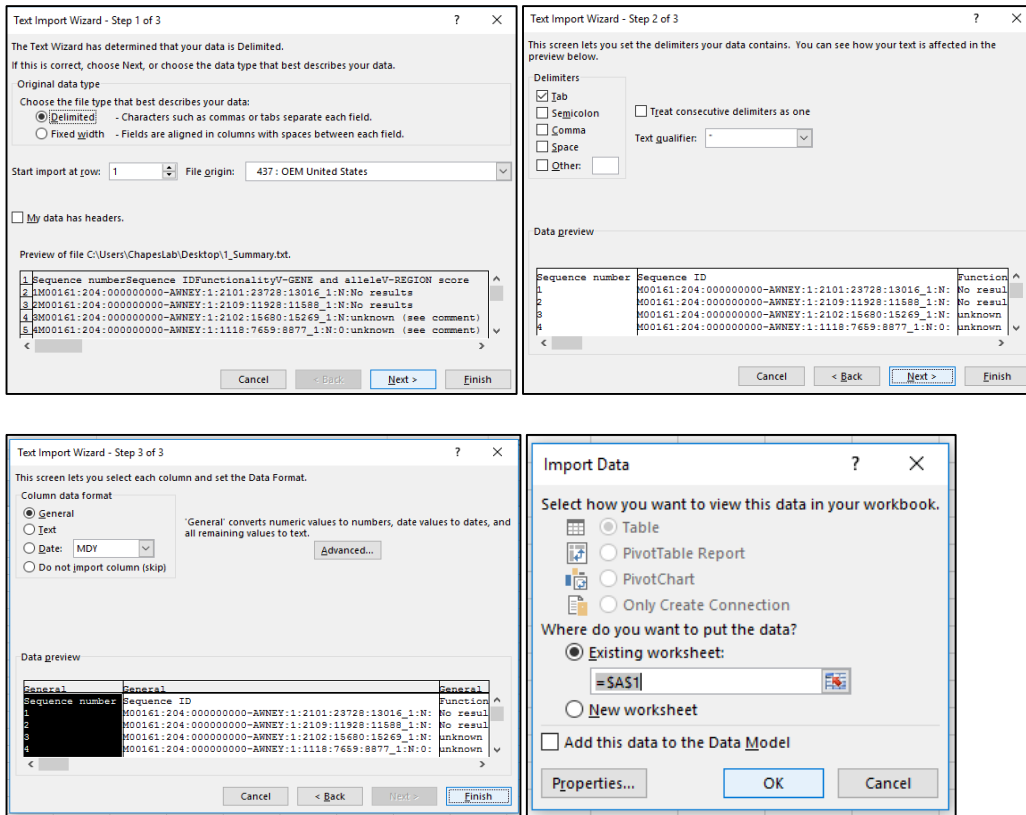




6. Select the 1\_Summary.txt to import into the first tab and select “Import”.



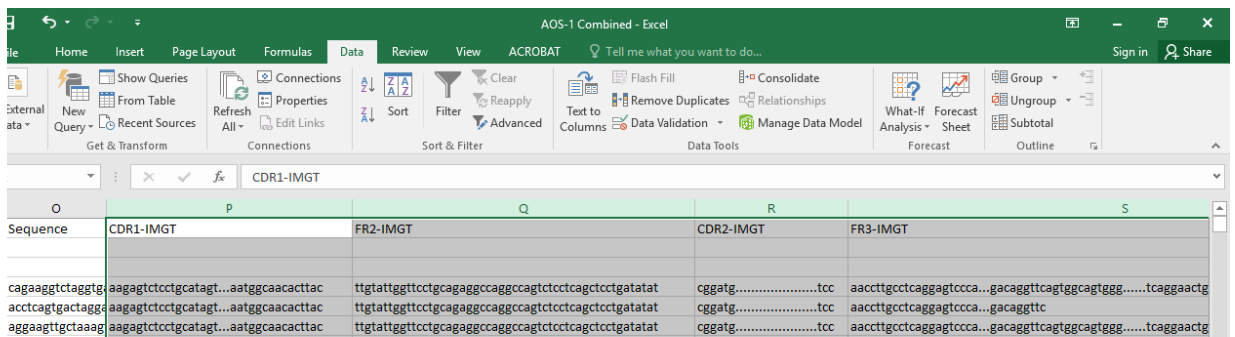
7. Complete the import by selecting the options shown in the screenshot below.



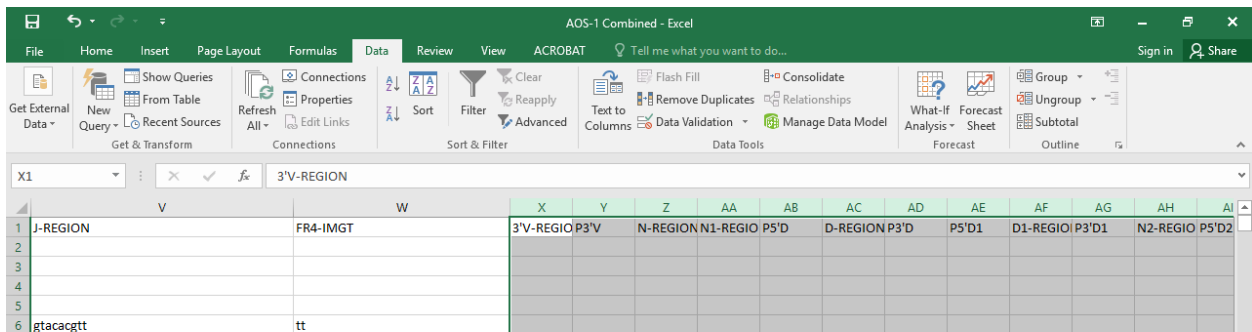
8. Delete the necessary columns as specified on the IMGT output field list (Page 1).

9. Create a new tab and import the IMGT output file “2\_IMGT-gapped-nt-sequences” using the import directions from steps 5-8.

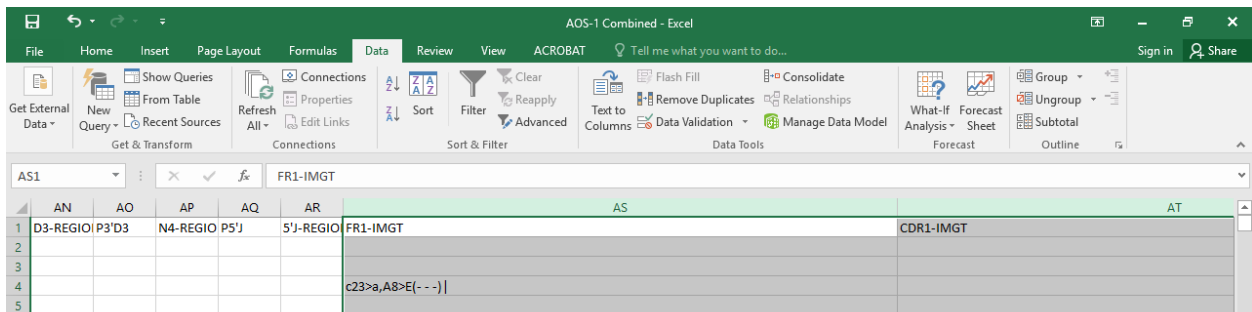
10. Highlight the necessary columns as specified on the IMGT output field list, copy the columns and then paste the columns into the end of the first tab (which contains data from “1\_Summary.txt”).



11. After necessary fields have been copied and transferred to the first tab, delete the second tab.
12. Create a new tab and import the IMG\_T output file “6\_Junction” using the import directions from steps 5-8.
13. Highlight the necessary columns as specified on the IMG\_T output field list, copy the columns and then paste the columns into the end of the first tab (which contains data from “1\_Summary.txt” and “2\_IMG\_T-gapped-nt-sequences.txt”).

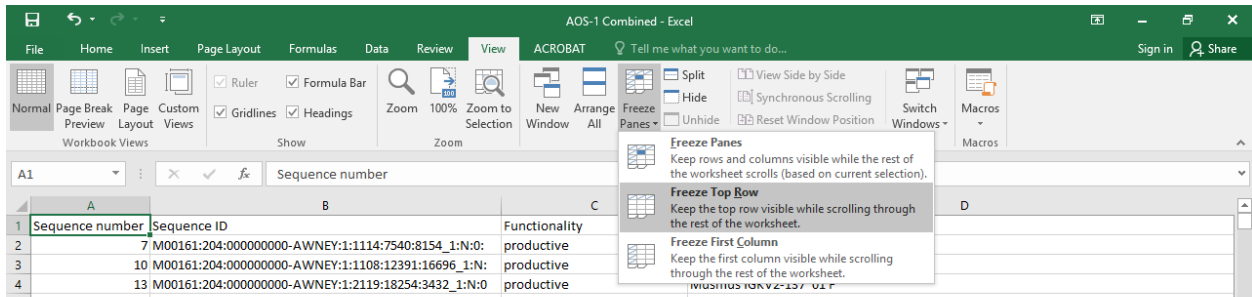


14. After necessary fields have been copied and transferred to the first tab, delete the second tab.
15. Create a new tab and import the IMG\_T output file “7\_V-REGION-mutation-and-AA-change-table” using the import directions from steps 4-6.
16. Highlight the necessary columns as specified on the IMG\_T output field list, copy the columns and then paste the columns into the end of the first tab (which contains data from “1\_Summary.txt”, “2\_IMG\_T-gapped-nt-sequences.txt” and “6\_Junction”).



17. After necessary fields have been copied and transferred to the first tab, delete the second tab.
18. Rename the first tab “Combined” if only one IMG T submission was necessary (part 1 of 1) and skip to “Duplicate Removal”. If multiple IMG T submissions were necessary, name the tab “Combined Part 1” and proceed to the following steps.
19. Remove all previous IMG T output files from the desktop. This will prevent importing data from the same submission part multiple times.
20. Repeat steps 2-19 using the next IMG T submission part (i.e. AOS-1 Kappa Part 2 of 2). For ease of combining parts in future steps it is important to delete and add necessary columns in the same order on the combined tab for each submission part.
21. Continue process until all IMG T submission parts have been condensed into a single tab each.
22. Within each “part” tab, find and replace “ see comment” (note space before phrase) in the functionality column.
23. Within each “part” tab select all of the data, select the sort button on the data tab. Sort by Functionality from A-Z.
24. Within each “part” tab, delete all rows that contain “No results” under functionality. This step is done in each tab individually because in some instances there are too many rows to combine the data from all tabs prior to this cleaning step.
25. To combine all IMG T submission parts, select the “Combined Part 1” and ensure that row 1 is the top row in view on the tab. Select the “View”, “Freeze Panes” and “Freeze Top Row”. This will allow for the addition of the additional parts to the bottom of the

“Combined Part 1” tab while allowing for comparison of row headings for quality assurance.

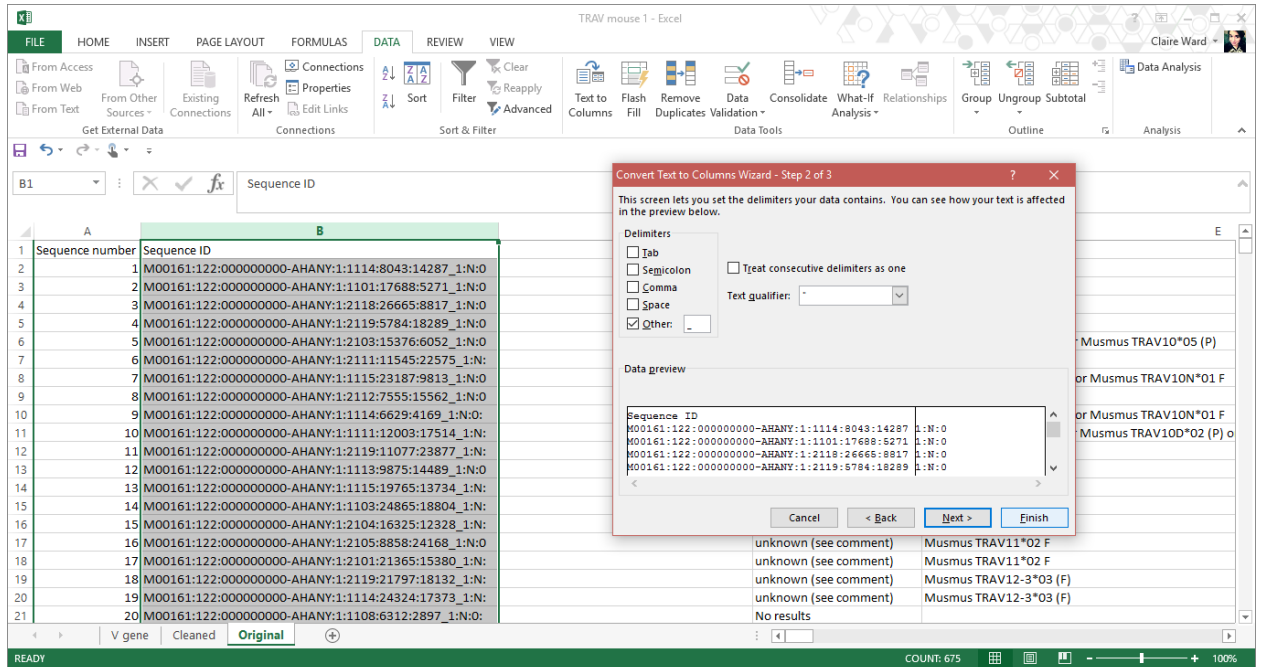


26. On the “Combined Part 2” tab, select all rows in the spreadsheet and copy the rows.
27. Select the “Combined Part 1” tab and paste the rows into the bottom of the spreadsheet.
28. Compare the column headings that were transferred from the “Combined Part 2” tab by scrolling to align the headings from both tabs.
29. Once all columns have been confirmed to align, delete the heading row for the copied part 2 data.
30. Repeat for additional combined part tabs if necessary, adding to “Combined Part 1” each time.
31. Save a version of the combined files for future reference and make a copy of the file to continue cleaning the IMG\_T output files.

## Appendix B.5 Duplicate read removal

1. In a copy of the combined IMG\_T output files, insert a blank column after the original “Sequence ID” column.

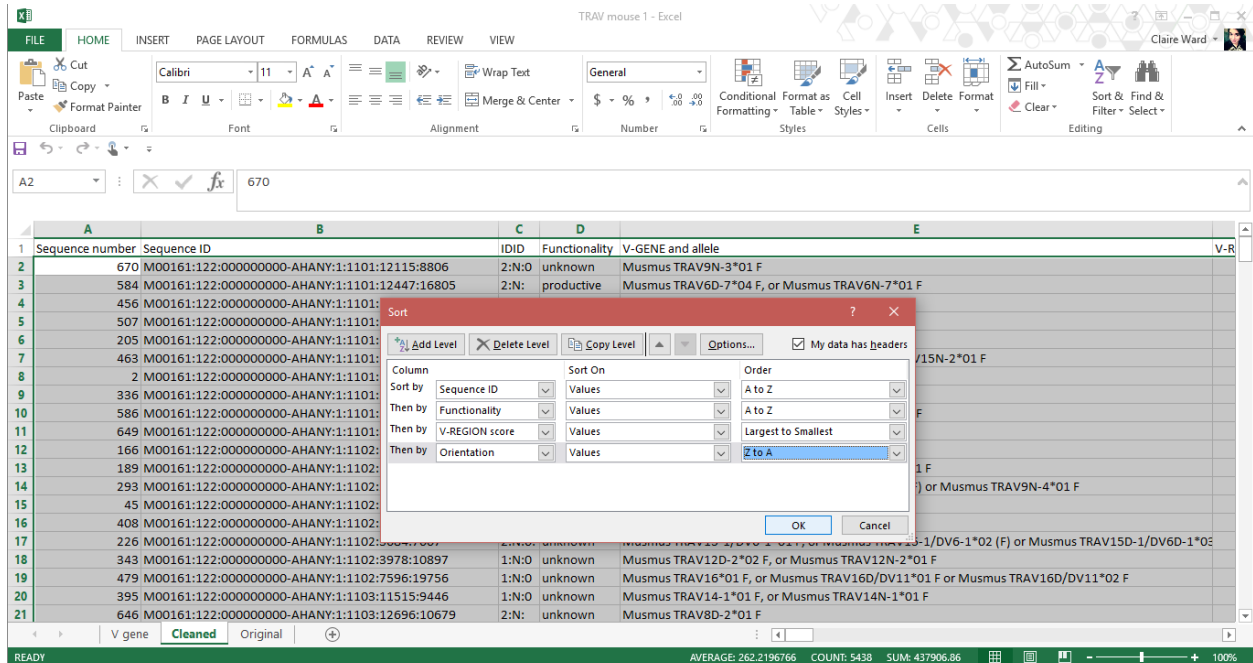
- Select the “Sequence ID” tab delineate based on the underscore character “\_”. This will all the end identifier from the copied sequence ID to the previously blank column D. This step will allow you to sort by the sequence ID in column D without those end identifiers which prevent like sequences from being combined. The new column is labeled “IDID” in the example shown below.



A	B	C	D	E
1	Sequence number	Sequence ID	IDID	Functionality
2	670	M00161:122:000000000-AHANY:1:1101:12115:8806	2:N:0	unknown
3	584	M00161:122:000000000-AHANY:1:1101:12447:16805	2:N:	productive
4	456	M00161:122:000000000-AHANY:1:1101:15208:14512	2:N:	unknown
5	507	M00161:122:000000000-AHANY:1:1101:15992:16513	1:N:	unknown
6	205	M00161:122:000000000-AHANY:1:1101:16962:3638	2:N:0	unknown
7	463	M00161:122:000000000-AHANY:1:1101:17431:6798	1:N:0	unknown
8	3	M00161:122:000000000-AHANY:1:1101:17689:5271	1:N:0	unknown

- Ensure that all rows with “no results” under the functionality column have been removed. This is an important step as duplicates will be removed after being sorted in alphabetical order. Any remaining “no results” will result in the incorrect duplicate values being removed. At this point it is optional to remove reads of “unproductive” functionality as they will sort last in the duplicate removal process and can be filtered out later.

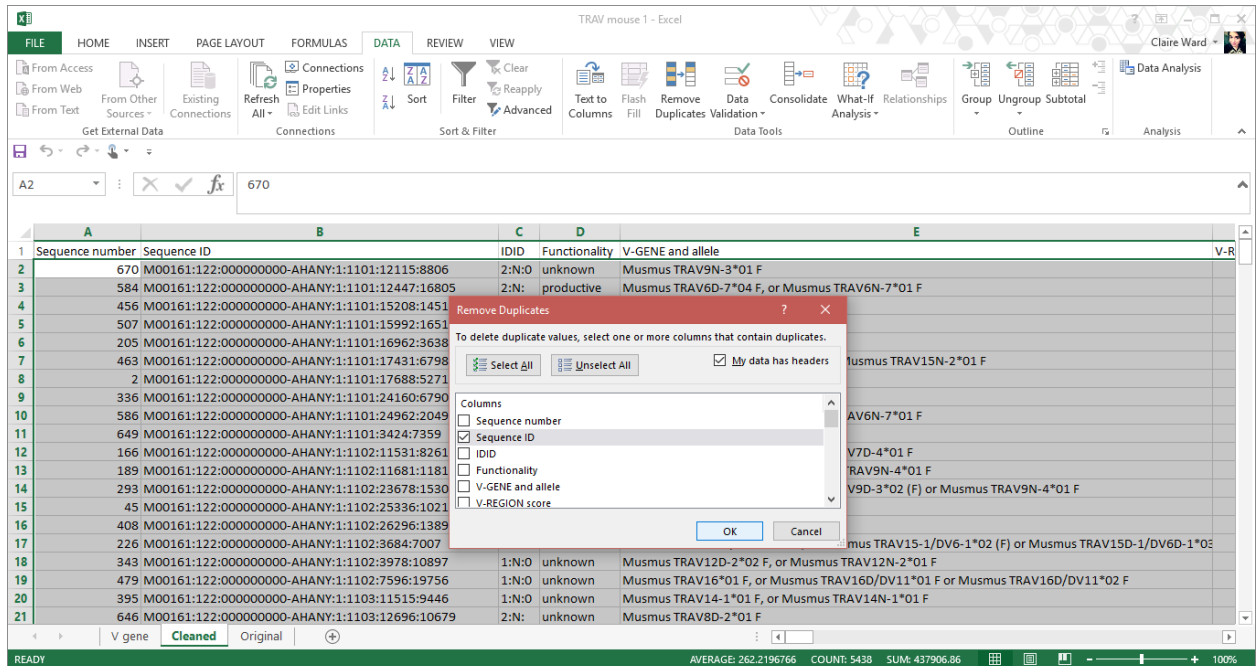
- Next, select all rows and columns for a “custom sort” (home tab, under sort and filter) and add levels to reflect the order as shown in the screen shot below.



The rationale for selecting these sorting columns (and sorting orders) is as follows:

- Sequence ID: This will group like IDs together to easily assess for duplicated. Keep the default A to Z ordering.
- Functionality: Between productive and unknown duplicate sequences, the productive sequence is preferred. Keep the sort order at A to Z because the productive, unknown and unproductive options are already listed A to Z.
- V-REGION score: The largest value among duplicate sequences is preferred. Sort this column largest to smallest to ensure selection of the highest V-REGION score.
- Orientation, sort this column Z to A. The orientations are either positive and negative. Sorting Z to A to select the positive orientation. This component of the sort was arbitrarily chosen.

- Select all data and under the “Data” tab, select the option to remove duplicates. A dialogue box will appear. Unselect all and then only select “Sequence ID”:



- A box will report the number of duplicates that were removed and how many unique sequences remain.

**Note:** CDR are identified by conserved motifs which include a C-xx-W motif for IgH and a C-xx-F motif for both Igκ and Igλ. Sequencing reads that do not fit these conserved motifs are included in the IMGT outputs as functionally productive antibodies. For this reason, the functionality of productive reads with non-conventional CDR3 should be re-designated as “unknown”. The procedure for this modification to the IMGT output is as follows.

### *Identification of non-conventional CDR3*

- Locate the column titles “AA JUNCTION”, insert 3 columns to the left of that column.
- Perform a find and replace in “AA Junction” for “(see functionality comment)” and replace with nothing. Note that a space is included before the phrase.



3. Label the first inserted column “Left” in row 1 and enter the following formula in row 2: =left([select “AA Junction” row 2 cell],1). This will select the leftmost character of “AA Junction”.
4. Label the second inserted column “Right” in row 1 and enter the following formula in row 2: =Right([select “AA Junction” row 2 cell],1). This will select the right-most character of “AA Junction”.
5. Label the third inserted column as “Conventional” and enter the following formula in row 2: =if(and([select “Left” row 2 cell]=”C”, [select “Right” row 2 cell]=”F”),1,0)
6. Highlight row 2 for the three new columns and fill down the formulas by double clicking the plus sign that appears when you hover over the lower right corner of the highlighted cells.
7. The “Conventional” column should now include binary coding (1=yes) for all “AA Junction” that fit the C-xx-F.

### ***Reassigning Functionality***

8. Insert a column to the right of the “Functionality” column.
9. Copy the contents of the “Functionality” column and paste them into the new column.
10. Change the column header of the new column from “Functionality” to “Updated Functionality”
11. Hide all columns that are located between “Updated Functionality” and “Conventional” by highlighting the intervening columns, right clicking and selecting “Hide”. This will make the next step easier to see.
12. Highlight the entire spreadsheet and perform a custom sort with the following levels: 1) “Functionality” (A-Z) and 2) “Conventional” (smallest to largest).

13. Go to the “Updated Functionality” column and change all “productive” cells to “unknown” where a “0” is found under the “Conventional” column.
14. Highlight the “Updated Functionality” and “Conventional” columns, right click and select ”Unhide” to view all data again.
15. The “Updated Functionality” column should be used for all repertoire assessments. V genes will contain reads of unknown and productive functionality, which all other assessments such as J and CDR3 will contain only productive reads.
16. Retain a version of the excel spreadsheet as a reference, name the spreadsheet “removed duplicates”.

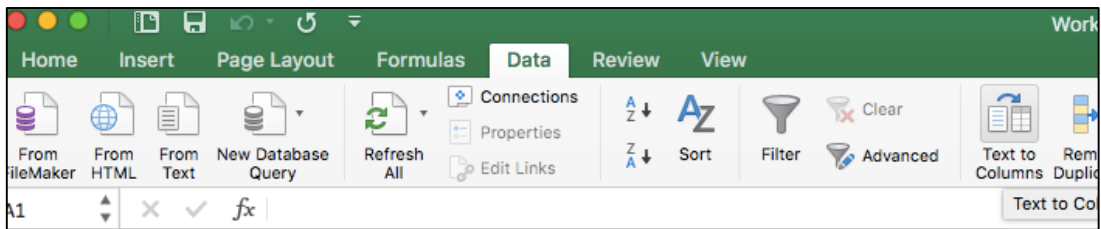
## **Appendix B.6 Assessment of V Gene Segment Usage**

### ***Cleaning the V-gene segment column***

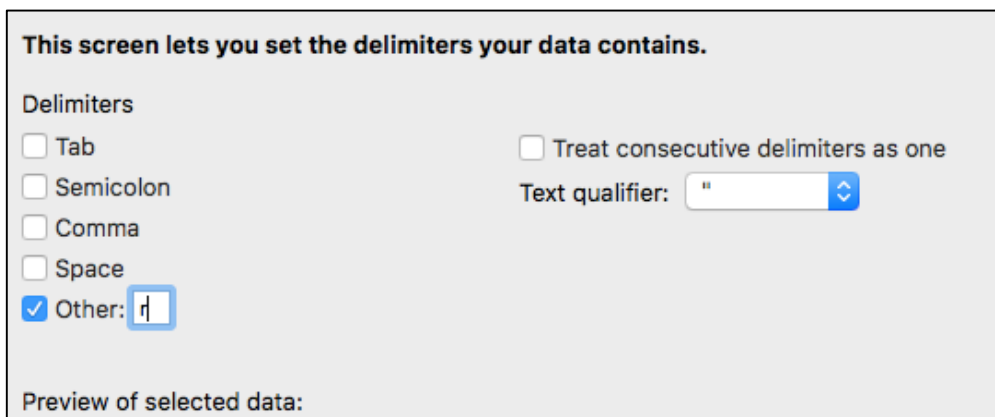
17. Create a copy of the removed duplicates spreadsheet and name “V gene”.
18. Insert a new column on the far left-hand side. Name the column “Sort ID”.
19. In the second cell of the new “Sort ID” column, type 1. In the third cell, type 2. In the fourth cell, type 3. Select the three numbered cells and fill down the formatting to the end of the spreadsheet.
20. Clean the “V-gene and allele” column by performing a find and replace (control + H) on the following terms using a blank replace field:
  - a. (see comment)
  - b. musmus
  - c. ORF
  - d. F

- e. ,
- f. o
- g. ()
- h. []
- i. space bar

21. Additionally, some non-*Mus musculus* reads may be present and are seen as musspr. Because the column will be tab delineated on “r”, search for any “musspr” and replace with “mussp”.
22. Add roughly 15 columns to the right of the “V-Gene and allele” column (enough to account for the possible duplicates).
23. Select the V gene and allele row and then under the data button, select “Text to columns”



24. In the window that opens, select “Delimited” and “next”. Then check the “other” box and type r in the open field.



25. Your data will now be sorted into many columns. If you have not added enough empty columns for the data to be parsed, you will see a dialog box that will ask you if you will allow the data to replace data from other cells. Select no or cancel and add more columns, then repeat steps 6 and 7.
26. Go to the last column and highlight the column from the top, view in the lower right hand bar whether values are present in the cells. If no value appears, delete the column and repeat until a value appears.
27. Note the last column in which a delineated V gene segment appears (example column K).
28. Select all data and perform a custom sort in order starting with the last V-gene segment column and then add secondary sort fields in reverse alphabetical order until you reach the original V gene and allele column. (Ex. Sort by: K, J, I, H, G, V gene and allele)
29. The sorting result is a reverse pyramid of the reads with the most duplicate gene segments appearing at the top.
30. Select all rows that have a duplicate read and copy. Paste the rows at the bottom on the spreadsheet.
31. Select the first of the newly copied cells in the “V gene and allele” column, highlight from that cell to the bottom of the column and delete, shift cells to the left.
32. From that section, copy the top row with duplicate reads to the bottom of the spreadsheet. Paste the newly copied cells at the bottom of the spreadsheet and repeat step 16.
33. Repeat this process until no more duplicate cells are seen.
34. Now align the data to the right of the “V gene and allele” column by selecting the lowest duplicate allele cell and pressing control+shift and the up arrow until the top of the column

is selected. Delete the selection and shift cells to the left. Continue this process until the “V region score” column is aligned to the right of the “V gene and allele” column for all cells.

### ***Identifying C57BL/6 Alleles and Reads with multiple C57BL/6 options***

**Note: Not all IMGT outputs will provide gene segment alleles that are specific to C57BL/6 mice. The VLOOKUP function will be used to compare the cell entry for “V gene segment and allele” to a list of C57BL/6 genes. “NA” will be returned if the contents of the cell are not found on the list.**

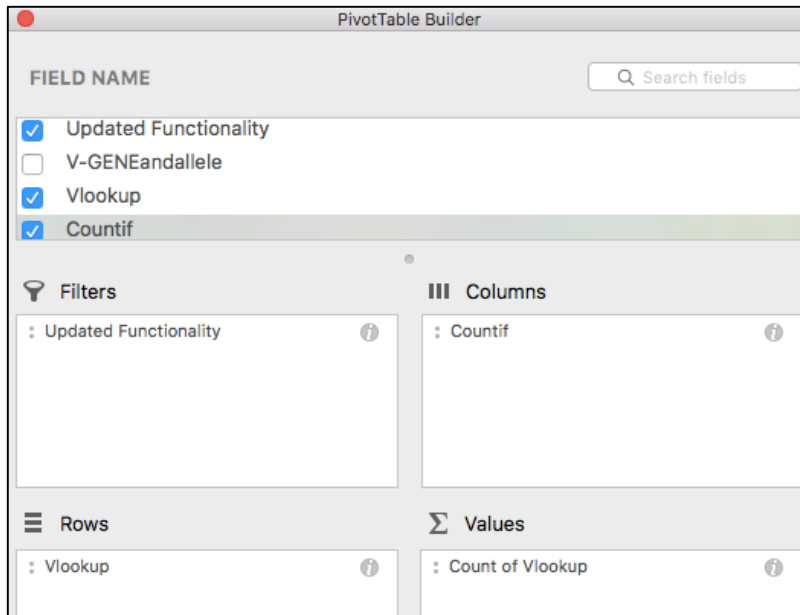
1. Create a new tab that named “B6 reference” and copy in the C57BL/6 reference list for IGKV. This file is a word document named “IGKV B6 Reference” located on the lab computer under: C:\Users\ChapesLab\Documents\Claire\IGKV FASTA. Currently, pseudogenes have been removed from this list.
2. Insert a new column to the left of the “V gene and allele” column and title this column “Vlookup”
3. For the first read in this column set up the following formula =VLOOKUP(Value,List,1,FALSE). The value will be the “V gene and allele” cell in that row. The list will be the B6 gene segments on the separate reference tab that was just created. The coding 1 and FALSE ensure that the contents of the cell and reference list have to be an exact match, and that “NA” will be reported in the event that the value does not match a reference entry.
4. Fill down this formula to all cells in the column.
5. Select all data and custom sort on the “Vlookup” column A-Z.

6. Scroll down to where the “NA” output start in the “Vlookup” column and then select the “Sort ID” cell from that row. Highlight all cells below that in the “Sort ID” row and delete the contents. This step is done to remove all of the sort IDs from non-C57BL/6 read entries.
7. Insert a new column to the left of the “Vlookup” column and title the column “countif”
8. In the first cell of this column, enter the formula =countif(range,value). The range will be the entire “Sort ID” column and the value will be the “Sort ID” cell in that row. Fill down this formula in the “countif” column. This will return the number of times the value appears in the column (anything that had multiple C57BL/6 alleles will have a value greater than 1).

### ***Determining percent abundance of V gene segment use***

**Note: In this step a pivot table is used to select for different variables of interest. Pivot tables will take multiple cells with the same value and generate a tally.**

1. Highlight the “Updated Functionality” column through the “count if” column.
2. On the insert tab, select “PivotTable”.
3. The range highlighted will appear as the range selected. Select the option to insert the table in a new spreadsheet. Select Ok.



4. Drag the “countif” variable into the “columns” square, the “Vlookup” variable into the “rows” and “sum values” squares, and the “Updated Functionality” variable filter into the filters square.
5. Select the drop down arrow for the filter and uncheck the “Unproductive” and “Blank” boxes. This will show only the reads that were identified as having productive or unknown functionality.
6. Select the drop down arrow for the columns and uncheck the all “countif” options and recheck “1” and “2”. This will show only the reads that had only one or two C57BL/6 allele(s) assigned.
7. Finally, Select the row drop down arrow (VLOOKUP) and uncheck the “NA” and “Blank” boxes. This will remove all of the non-C57BL/6 V genes that were grouped together using the VLOOKUP function.

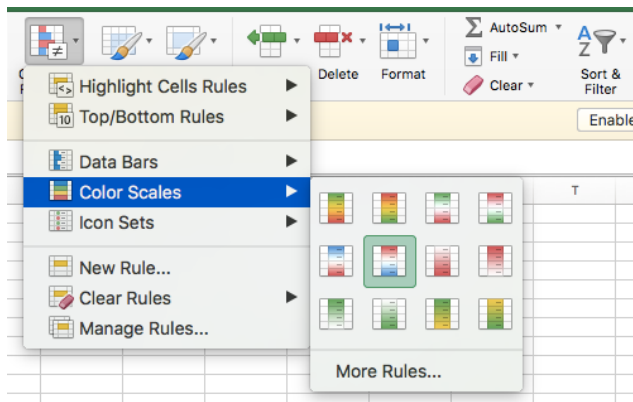
Updated Functionality (Multiple Items) ▾				
Count of Vlookup				
Column Labels ▾				
Row Labels ▾	1	2	Grand Total	
IGKV1-110*01	1158	3	1161	
IGKV1-117*01	1962	4	1966	
IGKV1-122*01	115	2	117	
IGKV1-131*01	2	5	7	

8. Check the row and column variables to make sure that there are no duplicate values. If duplicates are present, there is likely a space that was not removed and the cells are counted as two different entries rather than being tallied together. If this happens, delete the pivot table and perform a control and replace of a spacebar in the column that the duplicates came from. Once corrected, remake the pivot table.
9. If the table contains no duplicates, copy the contents and paste into a new sheet.
10. Remove the totals from the bottom of each column.
11. Add a 0.5 weight to reads with a countif value of “2” by creating a new column that multiplies 0.5 by the values of the “2” column.
12. Sum the full read count (countif=1) and weighted partial column (countif=2) for each gene segment.
13. Calculate the percent abundance by selecting the summed full and partial read column in the following formula:  $=\text{value}/\text{sum}(\text{column}) * 100$ . Where value is the number of reads corresponding to the first gene segment row and the range is the sum of the entire column.
14. Fill down these formulas to obtain the percent abundances for all gene segments.

Row Labels	1	2	Grand Total	2*0.5	Sum	% Abundance
IGKV1-110*01	1158	3	1161	1.5	1159.5	4.69908815
IGKV1-117*01	1962	4	1966	2	1964	7.95947315
IGKV1-122*01	115	2	117	1	116	0.47011145
IGKV1-131*01	2	5	7	2.5	4.5	0.01823708
IGKV1-132*01	8		8	0	8	0.03242148
IGKV1-133*01	35	1	36	0.5	35.5	0.14387031



15. When comparing across multiple animals, copy the percent abundance into a separate spreadsheet and ensure that gene segment rows are aligned.
16. Separate columns can be added to indicate the rank of the gene segment within the animal by using the formula: `=RANK(cell,column,0)`, where “cell” is the percent abundance of the V-gene segment and “column” is the entire percent abundance column. “0” indicates that the value with the highest percent abundance will be ranked as 1.
17. A heatmap of gene segment usage can be generated by selecting percent abundance or rank of all animals and selecting color scales under conditional formatting.



## Appendix B.7 J gene segment usage

1. Create a copy of the removed duplicates spreadsheet and name “J gene”.
2. Sort the spreadsheet by “Updated Functionality” and remove all reads of “unknown” or “unproductive” functionality.
3. Perform a find and replace within the “J-gene and allele”, replacing the entry “less than six nucleotides are aligned” with “<6nt”.
4. Clean the “J-gene and allele” column by performing a find and replace (control + H) on the following terms using a blank replace field:
  - a. (see comment)

- b. musmus
  - c. ORF
  - d. F
  - e. ,
  - f. or
  - g. ()
  - h. []
  - i. space bar
5. The only J gene segments that are attributable to C57BL/6 mice are IGKJ1\*01, IGKJ2\*01, IGKJ4\*01, IGKJ5\*01. To retain these gene segments only, perform a find and replace on the following gene segments to be replaced with the word “Undetermined”:
- a. IGKJ1\*02
  - b. IGKJ2\*02
  - c. IGKJ2\*03
  - d. IGKJ3\*01
  - e. IGKJ4\*02
6. Sort the “J-gene and allele” column alphabetically. Clean cells with multiple gene segments by scrolling to the bottom of each C57BL/6 J-gene segment section within the column. For example, at the bottom of the IGKJ1\*01 section a few reads may contain:
- a. IGKJ1\*01IGKJ2\*02
  - b. IGKJ1\*01Undetermine
7. If multiple C57BL/6 J-gene segments are listed, replace the contents with “Undetermined”

8. If one C57BL/6 J-gene segment is listed with one or more “Undetermined”, retain the C57BL/6 J-gene segment only and remove “Undetermined” from the cell.
9. Repeat at the end of each C57BL/6 J-gene segment section.
10. Replace any cell that contains multiple “Undetermined” and no C57BL/6 J-gene segments with only one “Undetermined”.
11. Determine percent abundance for J gene segments by constructing a pivot table.
12. In contrast to the V-gene segment, all reads are productive and partial gene segments are not determined for J gene segments. For this reason, only select the “J gene and allele” column and select the pivot table option.
13. Drag the “J gene and allele” variable into both the row quadrant and the sum quadrant to generate a tally of reads for each J-gene segment.
14. If J-gene segments apart from: IGKJ1\*01, IGKJ2\*01, IGKJ4\*01, and IGKJ5\*01 appear, return to the “J gene and allele” column to remove those entries. This will happen if a non-C57BL/6 entry is not removed initially.
15. If duplicates of a C57BL/6 gene segment appear in the pivot table, return to the “J gene and allele” column check where a space was not removed during the find and replace step.
16. After the pivot table is constructed, copy the values into a separate tab and remove the total row.
18. Determine percent abundance with the following formula in a column next to the total J-gene segment read count tallies with the following formula:  $=\text{value}/\text{sum}(\text{column})*100$ . Where value is the number of reads corresponding to the first gene segment row.
19. When comparing across multiple animals, copy this data into a separate spreadsheet and ensure that gene segment rows are aligned.

20. Separate columns can be added to indicate the rank of the gene segment within the animal by using the formula: =RANK(cell,column,0), where “cell” is the percent abundance of the J-gene segment and “column” is the entire percent abundance column. “0” indicates that the value with the highest percent abundance will be ranked as 1.
21. A heatmap of gene segment usage can be generated by selecting percent abundance or rank of all animals and selecting color scales under conditional formatting.

### **Appendix B.8 Gene segment combination**

1. Create a copy of the cleaned “V gene” spreadsheet and name it “V-J combination”.
2. Sort the “Updated Functionality” column from A-Z and remove all “unknown” or “unproductive” rows.
3. Sort the “Countif” row lowest to highest and remove all rows that do not contain the number “1”. This analysis only uses “full” V-gene segment assignments.
4. Sort the “Vlookup” column A to Z and remove all entries that contain “N/A” (if not removed during “countif” removal step”. This analysis only uses C57BL/6 V-gene segments.
5. Clean the “J gene and allele” column by following the instructions used for the J gene spreadsheet. This step is repeated for the V-J combination analysis because we are interested in only V-J pairs that contain a full C57BL/6 V-gene, whereas in the J-gene segment analysis we are including all productive reads regardless of their V-gene status.
6. Once the “J gene and allele” column is cleaned, add a column next to the “J gene and allele” column and name it V-J combo.
7. To combine V- and J-gene segment entries, enter the formula: =(V-gene segment cell & “”& J-gene segment cell). Now both gene segments will appear for each row entry in a

single column separated by a space. This column contains data used for V-J linear regressions.

8. To combine V-gene segment families with their J-gene segment, insert a column next to the “Vlookup” column and copy the “Vlookup” values into the column. Name the column “V family”
9. Next, insert four new columns to the right of the “V family” column.
10. Tab delineate the “V family” column on the hyphen. The extra blank columns are added to ensure that no data to the right of the “V-family” column are overwritten during the tab delineation.
11. Keep only the column with V family entries (IGKV1, IGKV2, etc), deleted the four columns that were added prior to the tab delineation.
12. To combine V family and J-gene segment entries, enter the formula: =(V family cell & “”& J-gene segment cell) into a new column and label it “family combo”. Now both V family and he J-gene segment will appear for each row entry in a single column separated by a space. This column contains data used for bubble charts and Circos plots.

**Note: Individual pivot charts can be used to tally V-J combinations, however, the same V-J combinations may not be present within all animals. To easily align all V-J combinations between animals, non-tallied individual V-J combination values from all animals can be combined into a separate spreadsheet prior to constructive a pivot table.**

13. Create a new spreadsheet and name it “V-J Combined”.
14. Label the first column “Animal ID”, the second column “V-J combo” and the second column, and the third column “family combo”.
15. Copy the entries for “V-J combo” and “family combo” into the “V-J combined spreadsheet.

16. Before adding a new animal, ensure that the Animal ID was entered for the previous entry.

This can be done by typing the animal ID into the first three rows and selecting the fill down option to the last entry.

17. Once all animals have been entered, copy the “family combo” column and delineate on the space to create “V family” and “J Family” columns.

Animal	V-J Combo	Family Combo	V Family	J Family
M2	IGKV2-109*01 IGKJ1*01	IGKV2 IGKJ1	IGKV2	IGKJ1
M2	IGKV5-39*01 Undetermined	IGKV5 U	IGKV5	U
M2	IGKV1-110*01 IGKJ1*01	IGKV1 IGKJ1	IGKV1	IGKJ1
M2	IGKV4-57-1*01 IGKJ1*01	IGKV4 IGKJ1	IGKV4	IGKJ1
M2	IGKV4-61*01 IGKJ1*01	IGKV4 IGKJ1	IGKV4	IGKJ1
M2	IGKV4-61*01 IGKJ1*01	IGKV4 IGKJ1	IGKV4	IGKJ1
M2	IGKV4-61*01 IGKJ1*01	IGKV4 IGKJ1	IGKV4	IGKJ1
M2	IGKV4-61*01 IGKJ1*01	IGKV4 IGKJ1	IGKV4	IGKJ1

18. Highlight the five columns and create a pivot table.

19. To generate a table that tallies information for the linear regression, select the options displayed in the screenshot below:

The screenshot shows the PivotTable Builder interface. On the left, under 'Fields, Items, & Sets', the following fields are selected: Animal, V-J Combo, Family Combo, and V Family. The 'Columns' section contains 'Animal'. The 'Rows' section contains 'V-J Combo'. The 'Values' section contains 'Count of V-J Combo'. On the right, the resulting pivot table is displayed:

Count of V-J Combo	Column Labels			
Row Labels	M1	M2	M3	Grand Total
IGKV1-110*01 IGKJ1*01	165	195	161	521
IGKV1-110*01 IGKJ2*01	156	166	135	457
IGKV1-110*01 IGKJ4*01	56	60	54	170
IGKV1-110*01 IGKJ5*01	54	656	76	786
IGKV1-110*01 Undetermined	5	5	13	23
IGKV1-117*01 IGKJ1*01	218	219	189	626
IGKV1-117*01 IGKJ2*01	133	176	113	422
IGKV1-117*01 IGKJ4*01	57	86	43	186
IGKV1-117*01 IGKJ5*01	243	96	84	423
IGKV1-117*01 Undetermined	9	6	11	26
IGKV1-122*01 IGKJ1*01	6	7	6	19
IGKV1-122*01 IGKJ2*01	11	24	14	49
IGKV1-122*01 IGKJ4*01	1	4	3	8
IGKV1-122*01 IGKJ5*01	22		9	31
IGKV1-122*01 Undetermined		1		1
IGKV1-132*01 IGKJ1*01	2	1	1	4
IGKV1-132*01 IGKJ2*01			1	1
IGKV1-132*01 IGKJ4*01		1	2	3
IGKV1-132*01 IGKJ5*01	1	1	2	4

20. Determine the percent abundance of V-J gene segment combinations. These values can be used for pairwise linear regressions in GraphPad.

21. To generate a table that can be used for constructing a Circos plot, select the options displayed in the screenshot below:

The screenshot shows the PivotTable Builder interface on the left and the resulting PivotTable on the right. In the PivotTable Builder, the 'FIELD NAME' list includes 'Animal', 'V-J Combo', 'Family Combo', 'V Family', and 'J Family'. 'Animal', 'V Family', and 'J Family' are checked. 'Family Combo' is selected. The 'Filters' section contains 'Animal'. The 'Columns' section contains 'J Family'. The 'Rows' section contains 'V Family'. The 'Values' section contains 'Count of V Family'. The PivotTable on the right has 'Animal' as the filter, 'Count of V Family' as the column label, and 'IGKJ1' as the row label. The table shows counts for various IGK categories (IGKV1 to IGKV20) across five animal categories (IGKJ2, IGKJ4, IGKJ5, U) and a Grand Total.

Animal	(Multiple Items)					
Count of V Family	Column Labels					
Row Labels	IGKJ1	IGKJ2	IGKJ4	IGKJ5	U	Grand Total
IGKV1	1794	1322	518	1403	100	5137
IGKV10	1228	405	127	109	110	1979
IGKV11	7	22	1	13		43
IGKV12	1477	678	287	520	54	3016
IGKV13	116	109	35	70	7	337
IGKV14	282	863	297	202	17	1661
IGKV15	231	158	60	76	17	542
IGKV16	140	104	87	137	14	482
IGKV17	6	330	203	274	13	826
IGKV18	27			5	6	38
IGKV19	525	295	92	45	16	973
IGKV2	251	622	243	277	22	1415
IGKV20	2					2
IGKV3	1908	797	472	419	92	3688
IGKV4	938	1796	1815	2746	148	7443
IGKV5	621	537	576	1420	38	3192
IGKV6	756	1280	772	1202	92	4102
IGKV7	15	4		69		88
IGKV8	1009	812	328	957	48	3154
IGKV9	394	439	206	169	83	1291
<b>Grand Total</b>	<b>11727</b>	<b>10573</b>	<b>6119</b>	<b>10113</b>	<b>877</b>	<b>39409</b>

22. Determine the percent abundance of family combinations for circo plot formatting.

23. To generate a table that can be used for constructing a bubble chart, select the options displayed in the screenshot below:

The screenshot shows the PivotTable Builder interface on the left and the resulting PivotTable on the right. In the PivotTable Builder, the 'FIELD NAME' list includes 'Animal', 'V-J Combo', 'Family Combo', 'V Family', and 'J Family'. 'Animal', 'Family Combo', and 'V Family' are checked. 'Family Combo' is selected. The 'Filters' section is empty. The 'Columns' section contains 'Animal'. The 'Rows' section contains 'Family Combo'. The 'Values' section contains 'Count of Family Combo'. The PivotTable on the right has 'Animal' as the filter, 'Count of Family Combo' as the column label, and 'M1' as the row label. The table shows counts for various animal categories (IGKV1, IGKV10, IGKV11, IGKV12) across three animal categories (M1, M2, M3) and a Grand Total.

Count of Family Combo	Column Labels				
Row Labels	M1	M2	M3	Grand Total	
IGKV1 IGKJ1	567	652	575	1794	
IGKV1 IGKJ2	421	497	404	1322	
IGKV1 IGKJ4	147	227	144	518	
IGKV1 IGKJ5	376	797	230	1403	
IGKV1 U	25	34	41	100	
IGKV10 IGKJ1	217	609	402	1228	
IGKV10 IGKJ2	125	158	122	405	
IGKV10 IGKJ4	31	78	18	127	
IGKV10 IGKJ5	39	48	22	109	
IGKV10 U	34	38	38	110	
IGKV11 IGKJ1		5	2	7	
IGKV11 IGKJ2	5	5	12	22	
IGKV11 IGKJ4			1	1	
IGKV11 IGKJ5	6	5	2	13	
IGKV12 IGKJ1	342	772	363	1477	
IGKV12 IGKJ2	197	246	235	678	

24. Copy the contents of the table into a new tab.

25. Remove the total row from the bottom of the table.

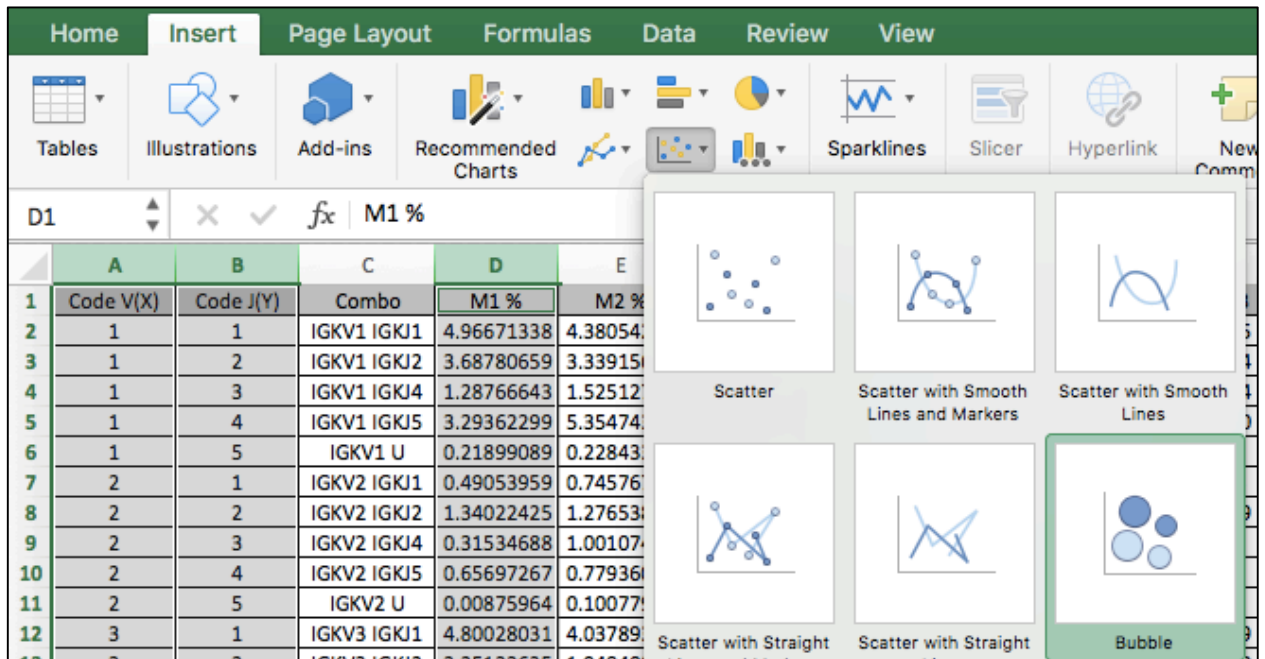
26. Determine the percent abundance of the family combos for M1, M2, M3 and Grant total columns. The grand total percent abundance column represents the average.

27. Insert two columns after the “Family combo” column and copy the “Family combo into the first new column.
28. Delineate the new “Family combo” column on the space into the remaining blank column and title the resulting two columns as “Code V” and “Code J”.
29. In the “Code V” and “Code J” columns, find “IGKV” and “IGKJ” and replace with nothing. The result are number codes for the bubble chart.
30. If all V-gene families (1-20) are present, no numeric gaps will appear in the bubble chart. If a family is missing, add the family to the column and enter “0” for the percent abundance of M1, M2, M3 and Grand Total.
31. The family number for “Code J must be changed to ensure no gaps are present and to allow undetermined to be included on the chart. For instance, J3 is a pseudo-gene. To remove that gap, J4 will be given the code “3” and J5 will be given the code “4”. Undetermined (or U) will be assigned the code “5”. Use the following coding key to find and replace within the “Code J” column:
  - a. Keep “1” as is
  - b. Keep “2” as is
  - c. Find “4”, replace with “3”
  - d. Find “5” replace with “4”
  - e. Find “Undetermined” or “U”, replace with “5”



Code V(X)	Code J(Y)	Combo	M1 %	M2 %	M3 %	Median %
1	1	IGKV1 IGKJ1	4.96671338	4.38054286	4.38629949	4.38629949
1	2	IGKV1 IGKJ2	3.68780659	3.33915614	3.08185216	3.33915614
1	3	IGKV1 IGKJ4	1.28766643	1.52512765	1.09848196	1.28766643
1	4	IGKV1 IGKJ5	3.29362299	5.35474335	1.7545198	3.29362299
1	5	IGKV1 U	0.21899089	0.22843322	0.31276222	0.22843322
2	1	IGKV2 IGKJ1	0.49053959	0.74576727	0.64078114	0.64078114
2	2	IGKV2 IGKJ2	1.34022425	1.27653856	2.1283088	1.34022425
2	3	IGKV2 IGKJ4	0.31534688	1.00107498	0.44244412	0.44244412
2	4	IGKV2 IGKJ5	0.65697267	0.77936039	0.65603784	0.65697267

32. Highlight “Code V” and “Code J” columns, in addition to the column for the desired bubble chart (For example, highlighting the “M1 %” column will generate a bubble chart for mouse 1 and highlighting “Median %” will generate a bubble chart that represents the average of all mice.



33. A bubble chart will appear, formatting of the bubble chart can be adjusted by right-clicking the data series and selecting “format data series”.

34. Under the “Series option” tab, the size of the bubble can be adjusted to reduce the level of overlap between coordinate.

35. In final formatting, axes titles should be removed and replaced with textboxes that clearly label the location of the gene family number, rather than the coded value.

## **Appendix B.9 CDR3 analyses**

1. Data for these analyses can be obtained directly from the completed “V-J combination” spreadsheet.
2. Maybe a copy of the “CDR-IMGT length” column and insert it to the right.
3. Under the new “CDR-IMGT length” column, separate out the CDR3 length by tab delineation. Insert XX blank columns to the right of the “CDR-IMGT length” column to prevent the overwriting of data from the tab delineation.
4. Delineate on the period “.” character.
5. Retain only the third CDR length from the delineated column. Label the column “CDR3 Length”.
6. The “AA Junction”, which contains the amino acid sequence of CDR3, does not need to be cleaned.
7. To assess CDR3 usage in multiple animals at once, create a new spreadsheet and name it “CDR3 analyses” combine the non-tallied values (individual rows) for “CDR3 length”, and “AA Junction” in a separate spreadsheet and along with the animal ID in a similar fashion to what was done for the “V-J combination” spreadsheet.
8. Highlight “Animal ID”, “CDR3 length” and “AA Junction” columns and insert a pivot table.
9. To construct a pivot table of CDR3 lengths, select the following options within the pivot table:

10. Do these other things to clean it up.
11. To construct a pivot table that aligns all CDR3 sequences, select the following output options:
12. Copy the resulting table into two new tabs.
13. In the first copied tab, determine the percent of repertoire for the CDR3 sequences in each animal as outlined in J-gene segment usage.
14. In the second copied tab, convert the CDR3 usage information a binary output for each animal. Sample coding is provided below for three animals (M1, M2, M3). Here, 1=present and 0=not found. Formulas can be entered into the first data row of separate columns and filled down for all CDR3 rows.
  - a. =if(CDR3 of M1>=0,1,0)
  - b. =if(CDR3 of M2>=0,1,0)
  - c. =if(CDR3 of M3>=0,1,0)
15. Using the binary coded CDR3 use column, determine the degree of overlap between animals. The example code for three animals (M1, M2, M3) below can be modified to accommodate more combinations. Enter these formulas into separate columns for the first data row and fill down the formulas.
  - a. CDR3 appears in M1 only: =if(and(M1=1, M2=0, M3=0),1,0)
  - b. CDR3 appears in M2 only: =if(and(M1=0, M2=1, M3=0),1,0)
  - c. CDR3 appears in M3 only: =if(and(M1=0, M2=0, M3=1),1,0)
  - d. CDR3 appears in M1 and M2: =if(and(M1=1, M2=1, M3=0),1,0)
  - e. CDR3 appears in M1 and M3: =if(and(M1=1, M2=0, M3=1),1,0)
  - f. CDR3 appears in M2 and M3: =if(and(M1=0, M2=1, M3=1),1,0)

g. CDR3 appears in M1, M2 and M3:  $==\text{if}(\text{and}(\text{M1}=1, \text{M2}=1, \text{M3}=1), 1, 0)$

16. The sum of each of these combination columns can be used to construct a Venn Diagram.