Inferring mechanisms of V(D)J recombination using statistical inference on high-throughput immune repertoire data

Magdalena L. Russell

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Frederick A. Matsen IV, Chair

Philip Bradley

Armita Nourmohammad

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Inferring mechanisms of V(D)J recombination using statistical inference on high-throughput immune repertoire data

Magdalena L. Russell

Chair of the Supervisory Committee:
Frederick A. Matsen IV
Department of Statistics

To appropriately defend against a wide array of pathogens, jawed vertebrates somatically generate highly diverse repertoires of B cell and T cell receptors through a random process called V(D)J recombination. Receptor diversity arises during recombination from the combinatorial assembly of V(D)J genes and the junctional deletion and insertion of nucleotides. While molecular experiments have established our understanding of V(D)J recombination *in vitro*, the processes underlying receptor generation *in vivo*, particularly in humans with intact recombination machinery, remain poorly characterized. This dissertation uses statistical inference on large immune receptor repertoire sequencing datasets to investigate the molecular mechanisms of V(D)J recombination in humans, with a focus on individual variability, nucleotide trimming, and the role of sequence microhomology. First, I identify genetic loci associated with modifying V(D)J recombination probabilities using genome-wide association inference and reveal individual differences in receptor generation. Next, I develop a probabilistic model of nucleotide trimming to infer how sequence-level features influence this process. Finally, I demonstrate that germline-encoded microhomology biases both trimming and ligation outcomes, providing mechanistic insights into its role in recombination. Together, these findings advance our understanding of how immune receptor diversity is generated and establish a foundation for future research on individual variability in immune responses.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Joining Erick Matsen's group has been the best decision of my academic career. I first met Erick in fall of 2019, when he taught several lectures for my favorite graduate course, Tools for Computational Biology. His enthusiasm for using math and computers to study biology stood out immediately, as did his extraordinary patience with students like me, who had very limited coding experience. Right away, I knew I wanted to work with him.

From that first class onward, Erick has been a remarkably supportive, thoughtful, and caring mentor. He approaches every interaction with intention, challenging me to grow into a rigorous and independent scientist while encouraging me to prioritize my well-being outside of work. Erick consistently models effective and generous collaboration, offering support and insightful feedback with an impressively quick turnaround. Perhaps most importantly, he has an incredible ability to recognize when I am struggling——often before I do——and provides the encouragement and affirmation I need to move forward. I feel incredibly lucky to have had such an exceptional mentor. While it is impossible to fully capture everything I have learned from Erick, I can only aspire to create a similarly supportive and intentional environment for my future trainees.

In addition to Erick, I am grateful to the wonderful members of the Matsen group, who have made my time here especially enjoyable. Mackenzie Johnson, Lena Colliene, and Kevin Sung (the "lunch bunch") made lunch and coffee breaks a source of daily entertainment and laughter. Hugh Haddox fostered community by organizing frequent gatherings and birthday

celebrations. Zorian Thornton has been a helpful resource for discussing graduate school and post-PhD aspirations. I collaborated with Assya Trofimov on a project [1] and learned so much from our frequent office whiteboard sessions. Jiansi Gao brightened my days with office chats and was a reliable ally in the battle against our unpredictable office temperature.

I have also been fortunate to work closely with several amazing individuals outside the Matsen group from whom I have learned so much. During my first year, Phil Bradley became a close collaborator and mentor. His enthusiasm for using math to solve data puzzles has been inspiring throughout my introduction to the field of computational immunology. I continue to be struck by his constant positivity, kindness, and remarkable ability to ask thoughtful, in-depth questions——even after long gaps between our meetings.

I am thankful for the chance to work with Dara Lehman and her lab during my Matsen group rotation. I learned so much from her ability to navigate interdisciplinary collaborations. Traveling to Kenya together to present our work at a conference was a highlight of my graduate school experience.

The flexibility of the MCB program allowed me to explore new areas and chart my own graduate school path, which was a transformative experience. I appreciate my past and present committee members——John Ray, Ellen Wijsman, Noah Simon, Phil Bradley, and Armita Nourmohammad——who have provided thoughtful feedback on my dissertation work and steady support as I navigated my post-PhD plans. I am fortunate to have been part of a wonderful MCB cohort and appreciate their friendship and support, especially from Cera Hassinan and Carrie Stine. I am also grateful to Maia Low and Denise Barnes at MCB and Ruby San Pedro with Fred Hutch Computational Biology, whose administrative expertise kept me on track.

I owe a tremendous acknowledgment to the mentors and teachers from earlier in my life who set me on a path to a PhD. My high school biology teacher, Paul Andersen, first sparked my excitement for biology. At Montana State University, Honors College Dean Ilse Mari Lee

became a pivotal mentor, providing unwavering encouragement and consistently guiding me toward opportunities that fostered my growth. She introduced me to Frances Lefcort, who shared her enthusiasm for research and guided me in defining my own interdisciplinary interests, even when they diverged from her own. In the Lefcort lab, I had the privilege of working closely with Sarah (Ohlen) Rogers, then a PhD student, who patiently taught me wet lab techniques, guided me through mistakes, and offered a supportive ear and fresh perspective whenever I needed it most.

Lastly, I want to thank my family and close friends, whose unwavering support has meant everything to me throughout this journey. My parents, Jim and Ronda, have been my greatest cheerleaders——(almost always) answering my calls on the first ring, celebrating even my smallest successes with unmatched enthusiasm, and providing endless affirmation during tough times. My brother, John, has taught me the importance of prioritizing happiness, and I've gained so much from his unique perspectives. My grandma, Gloria Lindemeier, has always been there for me, visiting me in Seattle and making my favorite foods (rhubarb sauce) whenever I come home. I am so grateful to have my family in my corner.

Will Dumm has been my greatest companion, always finding ways to make me laugh, offering thoughtful advice when I need it most, and meticulously planning our outdoor adventures——everything from small weekend hikes to (sometimes overly ambitious) long backpacking trips. His steady presence and support have been invaluable.

Don Dumm and Susan Payne have treated me like family, and I am so grateful for their constant encouragement, generosity, and frequent visits to Seattle. Our shared Katmai backpacking adventures with my family are my favorite memories of the past few years.

The Seattle-based Dumm family——Sue, Ryan, Mollyrose, Mitty, Ivy, and Lucie——has given me a wonderful sense of belonging in the area. They've welcomed me to countless dinners and gatherings, always remembering even the smallest details about my work and life. Mitty and Ivy, in particular, have involved me in more games of hide-and-seek than I

ever thought possible.

I am also thankful for my friends, especially Martha Krebill, Dave Biegel, and Courtney Linder, who have made this period of my life even more enjoyable. Clay Hunt has been an incredible example of determination; witnessing his PhD journey during my childhood inspired me more than he knows.

Finally, I want to acknowledge the friends and family members who are no longer here but whose impact on my life remains profound. Spending time with Tyler Dumm always left me with a sense of warmth; his thoughtfulness and curiosity were magnetic. Ralph Lindemeier and Nita Russell always made me feel loved and capable of achieving anything I set my mind to. I am so lucky to have crossed their paths.

Chapter 1

INTRODUCTION

Adaptive immunity relies on highly variable receptors on B and T cells to recognize and respond to specific pathogen-derived antigens. The diversity of these receptors is primarily generated through V(D)J recombination—a multi-step, stochastic process that assembles unique receptor sequences by joining V-, D-, and J-gene segments and modifying their junctions through nucleotide trimming and insertion.

This dissertation presents quantitative research aimed at inferring the mechanisms of V(D)J recombination and exploring how these processes vary across individuals. In Chapter 2, I investigate the influence of genetic background on V(D)J recombination probabilities using genome-wide association inference. Chapter 3 explores sequence-level factors that influence nucleotide trimming during V(D)J recombination using model-based statistical inference to provide new insights into the trimming mechanism. In Chapter 4, I examine the role of germline-encoded microhomology—short stretches of sequence homology—in biasing V(D)J recombination outcomes by developing the first probabilistic model that incorporates microhomology. Before presenting these studies, the following sections will summarize the relevant biological context.

1.1 Previous publication and co-authorship of dissertation content

This dissertation incorporates text and materials from the following previously published manuscripts [1–3], with contributions from all authors:

- Magdalena L Russell, Aisha Souquette, David M Levine, Stefan A Schattgen, E Kaitlynn Allen, Guillermina Kuan, Noah Simon, Angel Balmaseda, Aubree Gordon, Paul G Thomas, Frederick A Matsen, 4th, and Philip Bradley. Combining genotypes and T cell receptor distributions to infer genetic loci determining V(D)J recombination probabilities. *Elife*, 11, March 2022

- Magdalena L Russell, Noah Simon, Philip Bradley, and Frederick A Matsen, 4th. Statistical inference reveals the role of length, GC content, and local sequence in V(D)J nucleotide trimming. *Elife*, 12, May 2023

- Magdalena L Russell, Assya Trofimov, Philip Bradley, and Frederick A Matsen, 4th. Statistical analysis of repertoire data demonstrates the influence of microhomology in V(D)J recombination. *bioRxiv*, October 2024

1.2 The adaptive immune system

The adaptive immune system is defined by its ability to generate highly specific responses to a wide variety of pathogens. This specificity is driven by the vast diversity of antigen receptors expressed on its main cell types, T cells and B cells.

These cell types have distinct but complementary roles in the adaptive immune response. T cell receptors (TCRs) recognize pathogen-derived antigens presented by major histocompatibility complex (MHC) molecules on the surface of most cell types, allowing their corresponding T cells to detect the antigen-presenting cell. Once activated by MHC-antigen binding, naive T cells can differentiate into cytotoxic T cells, which directly kill infected cells, or helper T cells, which coordinate the immune response by activating other immune cells, including B cells. In contrast, B cell receptors (BCRs) can directly bind to pathogen-derived

antigens. Once activated, the corresponding B cells can secrete antibodies, the soluble form of their BCRs, which directly bind to and neutralize pathogens in extracellular spaces.

A large diversity of BCRs and TCRs is essential for recognizing a broad spectrum of potential pathogens. The processes by which these receptors are generated and selected will be summarized in the following section.

1.2.1 Overview of B and T cell receptor generation and selection

Both BCRs and TCRs are membrane-bound proteins that interact with antigens and share structural similarities. Both receptor types are built from protein chains forming immunoglobulin folds—polypeptide structures consisting of β-sheets linked by disulfide bonds. BCRs are composed of two identical heavy chains (encoded by genes within the *IGH* locus) and two identical light chains (encoded by genes within the *IGL* or *IGK* loci). Together, these chains form two antigen-binding regions located at the tips of the receptor. In contrast, TCRs are heterodimers made up of two distinct protein chains: the T cell receptor α (TCRα) and β (TCRβ) chains, which are encoded by genes from the *TRA* and *TRB* germline loci, respectively. A minority of T cells express an alternative receptor, also a heterodimer, consisting of γ and δ protein chains encoded by genes from the *TRG* and *TRD* loci. TCRs, unlike BCRs, have a single antigen-binding region.

The sequences encoding each receptor protein chain are generated through the V(D)J recombination process (discussed in detail in later sections). V(D)J recombination randomly selects V, D, and J gene segments from a large pool of germline-encoded genes, introduces nucleotide deletions and insertions at their junctions, and joins the segments together (Figure 1.1A). This mechanism can produce a vast diversity of receptor sequences.

Antigen specificity is determined by the complementarity-determining regions (CDRs) encoded within these recombined receptor sequences. The CDR1 and CDR2 regions are directly encoded by the V-gene segment, while the CDR3 region is assembled from parts of

the V-gene and the entirety of the D-gene (if present) and J-gene. The CDR3 region plays a particularly important role in antigen binding because it forms a flexible loop that extends to interact directly with the antigen. The junctional diversity introduced during recombination makes the CDR3 region the most variable among the CDRs, further enhancing its ability to mediate highly specific antigen interactions.

After V(D)J recombination generates TCRs and BCRs, B and T cells undergo a selection process in the bone marrow (for B cells) and thymus (for T cells). During selection, receptors are tested for proper expression and tolerance to self-antigens (Figure 1.1B). Because TCRs must recognize antigens presented by MHC molecules, they are also selected for their ability to recognize MHC. Cells that pass selection can then enter circulation, where they await exposure to their target antigens (Figure 1.1C). The collection of TCRs and BCRs formed through these processes constitutes the adaptive immune repertoire, which can be visualized analogously to a water pipeline, as illustrated in Figure 1.1D.

1.3 Sequencing-based approaches to studying adaptive immunity

Studying adaptive immunity is challenging due to the immense diversity, size, and dynamic nature of immune repertoires. High-throughput immune repertoire sequencing has emerged as a powerful approach for tackling these challenges by enabling the parallel sequencing of millions of receptor sequences [4–7]. These methods focus on capturing the complementarity-determining region 3 (CDR3) of BCRs and TCRs, the most variable antigen binding region.

Many software tools have been developed to infer the stepwise histories of V(D)J recombination events (i.e. gene choice, trimming, insertion, etc.) that produced each receptor sequence in these data [8–12]. These tools have enabled statistical analyses of receptor sequence distributions, which have provided valuable insights into global immune response dynamics [13, 14], as well as the mechanisms underlying TCR/BCR generation [11, 15] and

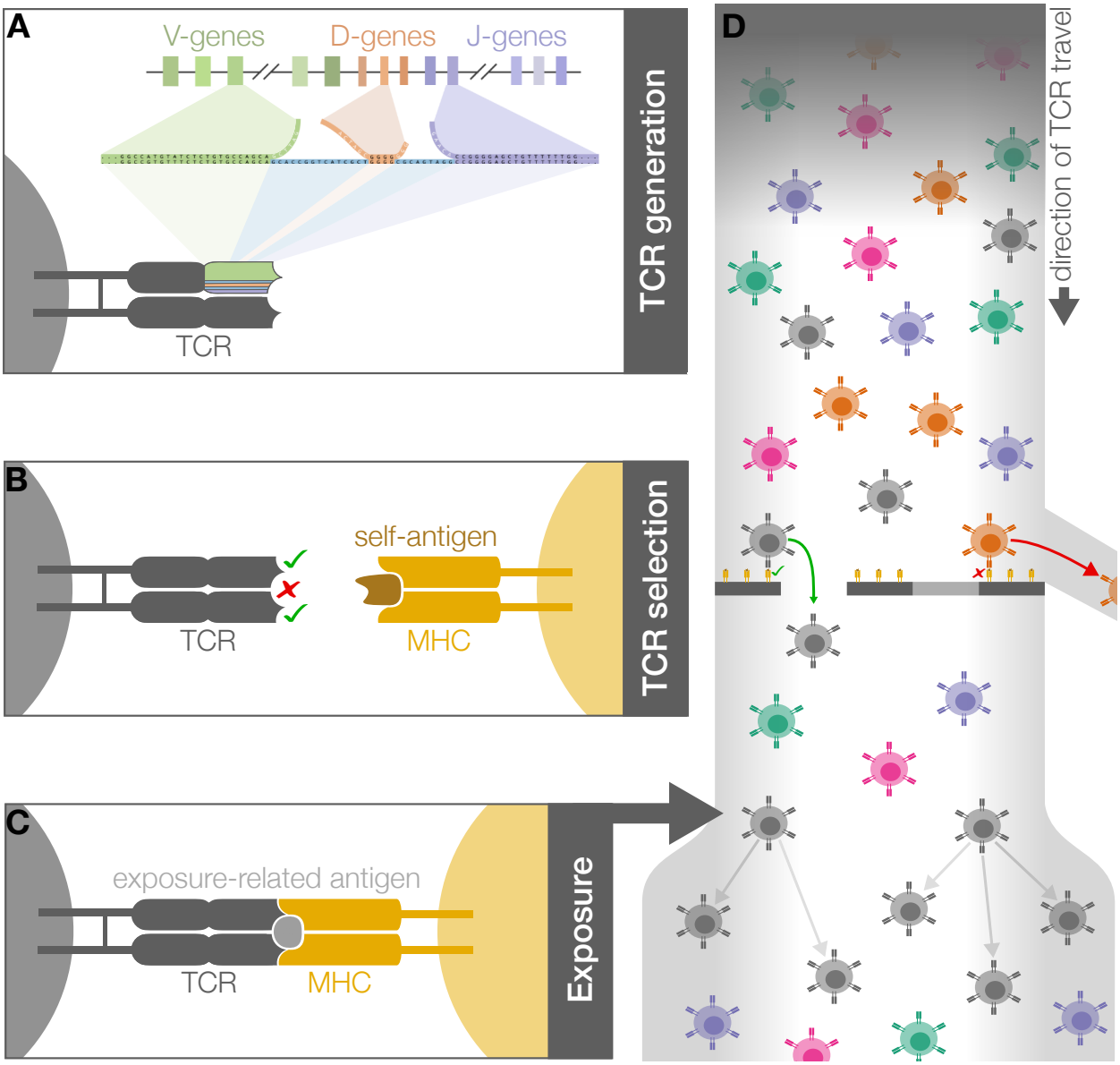

Figure 1.1: Schematic illustrating the processes of TCR generation, selection, and antigen exposure in forming a TCR repertoire. **(A)** TCRs (and BCRs) are generated through V(D)J recombination, where V-, D-, and J-genes are randomly selected from a large pool of germline-encoded genes, edited by nucleotide deletion and insertion at the junctions, and joined together. The CDR3 region is encoded by these junctions. This process can generate a vast diversity of receptor sequences. **(B)** Each TCR then undergoes selection for proper expression, MHC recognition, and self-tolerance; cells failing this selection process are removed. **(C)** Remaining cells circulate as part of the TCR repertoire, where they can encounter antigens from pathogen exposure, leading to rapid proliferation of the responding cells to mount an immune response. **(D)** The processes involved in TCR repertoire formation resemble a water pipeline: TCR generation acts as the water source, selection serves as a filter, and antigen exposure expands the pipeline through cell proliferation.

selection [16–18].

1.3.1 Productive and non-productive receptor sequences

TCR and BCR sequences can be classified into those that code for a complete, functional receptor (referred to as "productive" rearrangements) and those that do not (referred to as "non-productive" rearrangements). Non-productive sequences arise when the V(D)J recombination process produces a sequence that is either out-of-frame or contains a premature stop codon. Each developing T cell and B cell has two loci that can undergo the V(D)J recombination process. If the initial recombination attempt generates a non-productive sequence, a second recombination attempt can occur on the other chromosome, potentially producing a productive receptor. Non-productive rearrangements, though not expressed or functionally contributing to immune responses, can still persist in cells expressing a functional receptor and be sequenced as part of the repertoire.

Because non-productive sequences are not subject to functional selection, their recombination statistics offer a unique opportunity for studying the baseline V(D)J recombination process in the absence of selection pressures [11, 15, 19, 20]. In contrast, the recombination statistics of productive sequences reflect both the initial recombination process and subsequent selection-related effects.

1.3.2 Probabilistic modeling of immune repertoires

Probabilistic modeling has become an valuable tool for studying immune repertoires using high-throughput sequencing data, particularly for quantifying the generative and selective processes that shape repertoire composition. Models like IGoR [15] have advanced our understanding of V(D)J recombination by leveraging non-productive receptor sequences to learn statistical dependencies between recombination events——such as gene usage, nucleotide

trimming, and insertion——and quantify the probabilities of *generating* specific receptors. Extensions such as SONIA [17] and soNNia [18] build on IGoR by modeling selection pressures acting on productive sequences, thereby disentangling the probabilities of *generating* and *selecting* specific receptors.

More broadly, while studies in model organisms and in vitro systems have established a foundational understanding of receptor generation and selection, probabilistic modeling methods like these provide a robust framework for investigating these processes in vivo within a human context. In the next section, I will summarize our current understanding of the stepwise V(D)J recombination process, identify key knowledge gaps, and outline how I leverage modeling methods to address these gaps in this dissertation.

1.4 V(D)J recombination mechanism

As previously discussed, V(D)J recombination is a stochastic process that rearranges V, D, and J gene segments to generate unique receptor sequences that encode each protein chain of BCRs and TCRs. For TCR beta chains (encoded by the *TRB* locus) and BCR heavy chains (encoded by the *IGH* locus), recombination occurs in two stages: first, the D and J genes join, followed by the joining of a V gene to the D-J pair. In contrast, TCR alpha chains (encoded by the *TRA* locus) and BCR light chains (encoded by the *IGK* or *IGL* loci) undergo a single-stage process, joining a V gene directly to a J gene.

The step-by-step V(D)J recombination process is outlined below and summarized in Figure 1.2:

1. **Gene segment choice:** Gene segments are randomly chosen from a large pool of germline-encoded segments corresponding to each chain and gene type (e.g., *TRA* V-genes, *TRA* J-genes, *TRB* V-genes).

2. **Hairpin formation and opening:** The RAG complex removes the intervening chro-

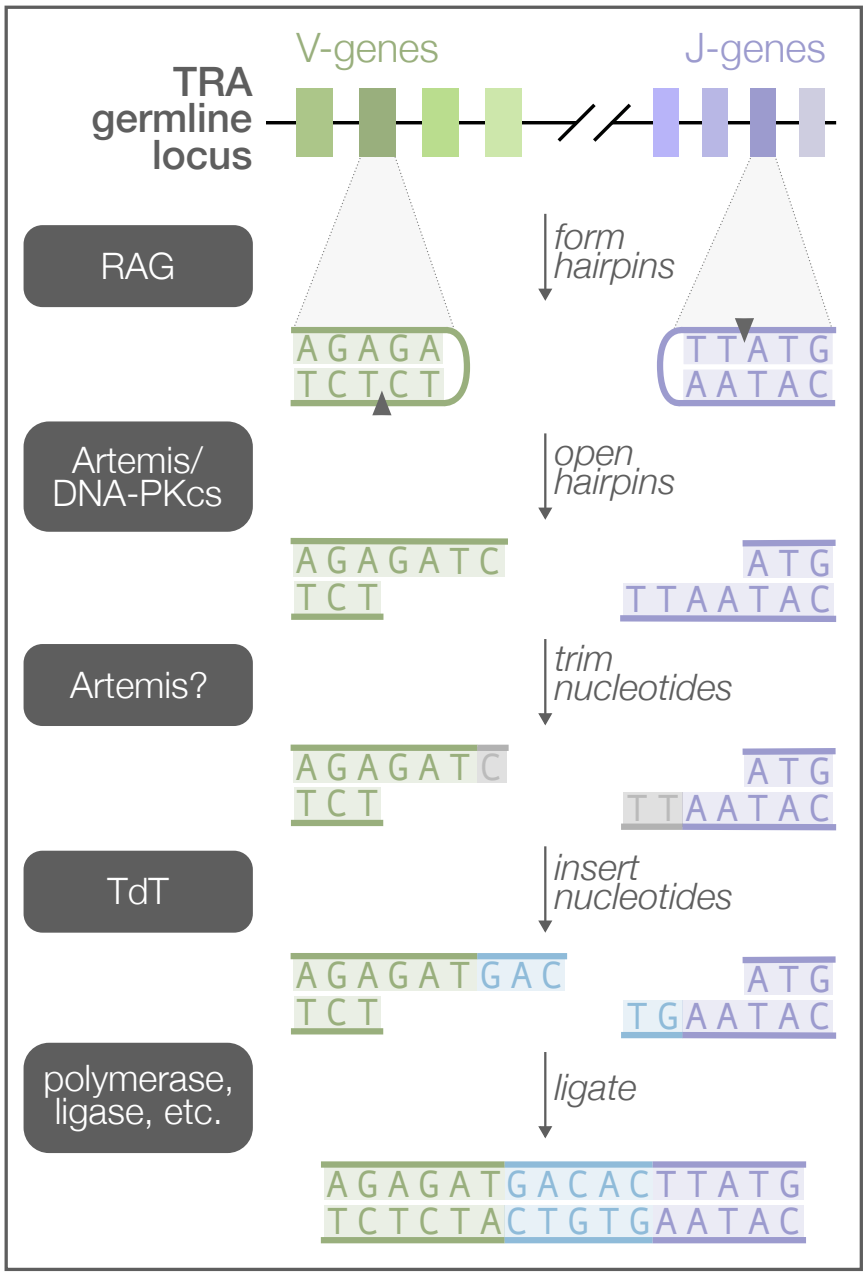

Figure 1.2: Schematic illustrating the steps of V(D)J recombination to join a *TRA* locus V-gene (green) with a J-gene (purple). First, the RAG complex randomly selects a specific V and J gene, removing the intervening chromosomal DNA and creating hairpin loops at the gene ends. Next, the Artemis:DNA-PKcs complex opens these hairpins, generating overhangs at each gene end. Nucleotides may then be trimmed from the gene ends (shown in gray; possibly by the Artemis protein) or inserted in a non-template-encoded manner (shown in blue) by the TdT protein. Remaining gaps between gene pairs are filled in by polymerases, and the genes are ligated to complete the joining process. Enzymes involved at each step are indicated in dark gray boxes.

mosomal DNA between the two gene segments and forms hairpin loops at the ends of each segment [21–23]. The Artemis-DNA-PKcs complex then asymmetrically nicks these hairpins, creating single-stranded overhangs that may include short palindromic sequences known as P-nucleotides, depending on the location of the nick [24–32].

3. **Nucleotide trimming and insertion:** Nucleotides can be trimmed from the end of each gene segment, possibly by the Artemis nuclease [25, 27, 30, 32–34], and non-templated nucleotides (N-insertions) can be added by terminal deoxynucleotidyl transferase (TdT) [35–37].

4. **Ligation**: Remaining gaps between the modified gene pairs are filled in by polymerases, and the genes are ligated together to complete the joining process. For TCRβ and BCR heavy chains, an additional joining step occurs to link the V segment with the D-J junction. For TCRα and BCR light chains, where only V and J segments are involved, recombination is complete after this single ligation step.

If recombination results in a productive sequence, the receptor is expressed on the cell surface and undergoes selection. In the following sections, I will describe several gaps in our understanding of the V(D)J recombination process and outline how this dissertation addresses them.

1.4.1 Genetic factors biasing V(D)J recombination

Immune repertoires are shaped by a complex interaction of genetically determined biases and immune exposures. Quantifying how genetic factors influence repertoire diversity is essential for understanding individual immune responses. However, the extent to which genetic background biases the V(D)J recombination process has yet to be fully explored.

Previous studies investigating the genetic and environmental determinants of TCR repertoire diversity and the V(D)J recombination process have been highly impactful de-

spite lacking high-throughput TCR repertoire sequencing data [38, 39] and/or high-resolution genotyping data [40, 41]. For instance, variation in the major histocompatibility complex (MHC) locus has been shown to bias TCR V(D)J gene usage [38, 39]. Additionally, biases in V(D)J gene usage, N-insertion lengths, and repertoire similarity in response to acute infection have been observed in monozygotic twins [40, 41].

Despite these findings, the lack of a comprehensive paired dataset has hindered genome-wide mapping of genetic influences on V(D)J recombination probabilities. In Chapter 2, I address this gap by quantifying the relationship between genetic background and V(D)J recombination probabilities, leveraging paired TCR immunosequencing and genotyping data from a large human cohort. This analysis reveals clear individual differences in receptor generation and provides new insights into the genetic determinants of immune repertoire diversity.

1.4.2 Nucleotide trimming mechanism

Nucleotide trimming is an essential diversity-generating step in V(D)J recombination. While the Artemis protein is often considered the main nuclease involved in V(D)J recombination, the precise mechanism by which it trims nucleotides remains uncertain.

Studies in model organisms and in vitro systems have shown that small variations in gene sequence can lead to large changes in the extent of nucleotide deletion [25–27, 30]. While these findings have provided valuable insights, the in vivo nucleotide trimming mechanism in humans is less understood. Statistical inference on high-throughput repertoire sequencing data now enables direct investigation of trimming processes within human systems.

To date, only one statistical analysis has linked sequence identity to trimming lengths [15]. This study demonstrated that a simple position-weight-matrix model could accurately predict trimming length distributions across various V-genes. However, the model's assumption that trimming is driven solely by sequence motifs limits its ability to explore alternative

mechanisms.

In Chapter 3, I investigate how sequence-level features influence nucleotide trimming probabilities using model-based statistical inference on high-throughput repertoire sequencing data. This approach provides new quantitative insights into how the Artemis nuclease may function to trim nucleotides during V(D)J recombination.

1.4.3 Microhomology in V(D)J recombination

In vitro studies have demonstrated that microhomology——short stretches of sequence homology between gene ends——can bias the V(D)J recombination process [26, 42–46]. However, the extent of microhomology's influence in vivo, particularly in humans, remains unclear. This gap in understanding holds both intrinsic scientific interest and practical implications: if microhomology biases recombination, it could influence the inference of V(D)J recombination annotations (i.e. the stepwise histories of events like gene choice, trimming, insertion, and ligation), potentially affecting the biological interpretation of immune receptor sequences used in downstream analyses.

As discussed earlier, existing probabilistic models of V(D)J recombination, such as IGoR [11, 15], have provided valuable insights into the mechanism of recombination. However, to our knowledge, no probabilistic models of V(D)J recombination incorporating microhomology have been developed.

In Chapter 4, I introduce the first probabilistic model of V(D)J recombination that incorporates microhomology to investigate how germline-encoded microhomology influences recombination outcomes in humans. This work deepens our understanding of microhomology's role in human V(D)J recombination and highlights the importance of accounting for microhomology-related effects during immune receptor sequence analysis and interpretation.

Chapter 2

COMBINING GENOTYPES AND T CELL RECEPTOR DISTRIBUTIONS TO INFER GENETIC LOCI DETERMINING V(D)J RECOMBINATION PROBABILITIES

Receptor proteins on the surfaces of T cells are an essential component of the cell-mediated adaptive immune response in humans. Cells throughout the body regularly present protein fragments, known as antigens, on cell-surface molecules called major histocompatibility complex (MHC). Each T cell expresses a randomly-generated T cell receptor (TCR) which can bind the MHC-bound peptide and, if necessary, initiate an immune response. As part of this immune response, a T cell will proliferate and subsequent clones of that T cell will inherit the same antigen-specific TCR. Over time, the collection of all TCRs in an individual (the TCR repertoire) will dynamically summarize their previous immune exposures [6].

To appropriately defend against a wide array of foreign pathogens, each individual has a highly diverse TCR repertoire. To generate diverse and functional TCRs, T cells combine a random generation process called V(D)J recombination with a selection process for proper expression and MHC recognition. Each TCR is composed of an α and a β protein chain which are both generated through V(D)J recombination. In the β chain, the recombination process proceeds by randomly choosing from a pool of V-gene, D-gene, and J-gene segments of the germline T cell receptor beta (TRB) locus over a series of

steps. First, the intervening chromosomal DNA between a randomly chosen D- and J-gene is removed to form a hairpin loop at the end of each gene [21–23]. Next, these hairpin loops are nicked open, often asymmetrically, by the Artemis-DNA-PKcs protein complex to create overhangs at the ends of each gene [24, 28, 29, 31, 32]. Depending on the location of the nick, the single-stranded overhang can contain short inverted repeats of gene terminal sequence known as P-nucleotides [25–27, 30]. From here, nucleotides may be deleted from the gene ends through an incompletely-understood mechanism suggested to involve Artemis [25, 27, 30, 32–34]. This nucleotide trimming can remove traces of P-nucleotides [26, 47]. Next, non-templated nucleotides, known as N-insertions, can be added between the gene segments by the enzyme terminal deoxynucleotidyl transferase (TdT) [35–37]. Once the nucleotide addition and deletion steps are completed, the gene segments are ligated together. The process is then repeated between this D-J junction and a random V-gene segment to generate a complete TCRβ protein chain. After the β chain has been generated, a similar α chain recombination proceeds, although without a D-gene, to complete the TCR. Following the generation process, each completed TCR undergoes a selection process in the thymus to limit autoreactivity and ensure its ability to correctly bind peptide antigens presented on a specific MHC molecule [48, 49].

TCR repertoires vary between individuals and are a complicated tangle of genetically determined biases and immune exposures. Disentangling these factors is essential for understanding how our diverse repertoires support a powerful immune response. Previous efforts to unravel the genetic and environmental determinants governing TCR repertoire diversity have been highly impactful despite lacking high-throughput TCR repertoire sequencing data [38, 39] and/or high-resolution genotype data [39, 40, 50, 51]. For example, it has been shown that variation in the MHC locus biases TCR V(D)J gene usage [38, 39] and has been associated with clusters of shared receptors in response to Epstein-Barr virus epitope [52]. Other studies have reported biases in V(D)J gene usage [40, 41, 53–56], N-insertion

lengths [40], and repertoire similarity in response to acute infection [41, 54] for monozygotic twins. While this work clearly illustrates that genetic similarity implies TCR repertoire similarity, the extent to which specific variations are associated with V(D)J recombination probabilities has not been fully explored.

In this chapter, I directly address the question of how an individual's genetic background influences their V(D)J recombination probabilities using large human discovery and validation cohorts for which both TCR immunosequencing data [50, 52] and genotyping data [57] are available. With the goal of identifying statistically significant associations between single nucleotide polymorphisms (SNPs) and TCR repertoire features of interest using these novel, paired datasets, we treat analysis as a genome-wide association (GWAS) inference with many outcomes. Our results suggest that MHC and *TRB* loci variations have an important role in determining the V(D)J gene usage profiles of each individual's repertoire. At the junctions, we demonstrate that variations in the genes encoding the Artemis protein and the TdT protein are associated with biasing V- and J-gene nucleotide deletion and V-D and D-J-junction N-insertion, respectively.

2.1 Results

2.1.1 Discovery cohort data description

We worked with paired SNP array and TCRβ-immunosequencing data representing 398 individuals and over 35 million SNPs and/or indels (Table 2.1). TCR sequences can be separated into those that code for a complete, full-length peptide sequence (which we will call "productive" rearrangements) and "non-productive" rearrangements that do not. Non-productive sequences can arise during TCR generation steps if the V- and J-genes are shifted into different reading frames or a premature stop codon is introduced in the junction region. A non-productive rearrangement can be sequenced as part of the repertoire when a recombina-

Table 2.1: Discovery cohort demographics.

		Count
Sex	Female	179
	Male	197
	Unknown	22
Age (in years)	< 10	12
	11-20	11
	21-30	48
	31-40	70
	41-50	103
	51-60	70
	> 60	22
	Unknown	62
Ancestry-informative PCA cluster (see Methods)	"African"-associated	8
	"Asian"-associated	23
	"Caucasian"-associated	322
	"Hispanic"-associated	30
	"Middle Eastern"-associated	5
	"Native American"-associated	10
CMV serostatus	Positive	171
	Negative	204
	Unknown	23
Total		398

tion event on one of a T cell's two chromosomes fails to create a functional receptor, followed by a successful recombination event on the other chromosome. Because these non-productive sequences do not generate proteins that participate in the T cell selection process within the thymus, they should not be subject to functional selection [15, 19]. As such, their recombination statistics should reflect only the V(D)J recombination generation process which occurs before the stage of thymic selection.

In the data cohort of 398 individuals, an average of 235,054 unique TCRβ-chain nucleotide sequences were sequenced per individual. Within each individual repertoire, roughly 18% of sequences were classified as "non-productive." Thus, we can analyze the productive and non-productive sequences separately to distinguish between TCR generation and selection effects within each TCR repertoire. Specifically, we inferred the associations between

Table 2.2: We inferred the associations between genome-wide variation and many different TCR repertoire features for productive and non-productive TCR sequences, separately. For each TCR repertoire feature, we considered the significance of associations using a Bonferroni-corrected threshold established to correct for each TCR feature subtype, the two TCR productivity types, and the total number of SNPs tested (described in detail in Methods).

Repertoire feature (significance threshold)	**Model type**	**Feature subtype**	**Sequence type**	**Significant association**
V(D)J gene usage (5.09×10^{-11})	simple	Each of 60 V-genes	Prod.	Yes, for 36 V-genes
			Non-prod.	Yes, for 26 V-genes
		Each of 2 D-genes	Prod.	Yes, for both D-genes
			Non-prod.	Yes, for both D-genes
		Each of 14 J-genes	Prod.	Yes, for 7 J-genes
			Non-prod.	Yes, for 8 J-genes
Amount of nucleotide trimming (9.68×10^{-10})	gene-conditioned	V-gene trimming	Prod.	Yes
			Non-prod.	Yes
		5' D-gene trimming	Prod.	No
			Non-prod.	No
		3' D-gene trimming	Prod.	No
			Non-prod.	No
		J-gene trimming	Prod.	Yes
			Non-prod.	Yes
Number of N-insertions (1.94×10^{-9})	simple	V-D-gene N-insertions	Prod.	No
			Non-prod.	Yes
		D-J-gene N-insertions	Prod.	No
			Non-prod.	Yes

genome-wide variation and V(D)J gene usage of each V-, D-, and J-gene, the extent of TCR nucleotide trimming, the number of TCR N-insertions, and the fraction of non-gene-trimmed TCRs containing P-nucleotides for both productive and non-productive sequences (Table 2.2).

2.1.2 TRB and MHC locus variation is associated with gene usage frequency

To quantify the effect of SNPs on the expression of various V-, D-, and J-genes during V(D)J recombination, we designed a fixed effects model to assess the relationship between

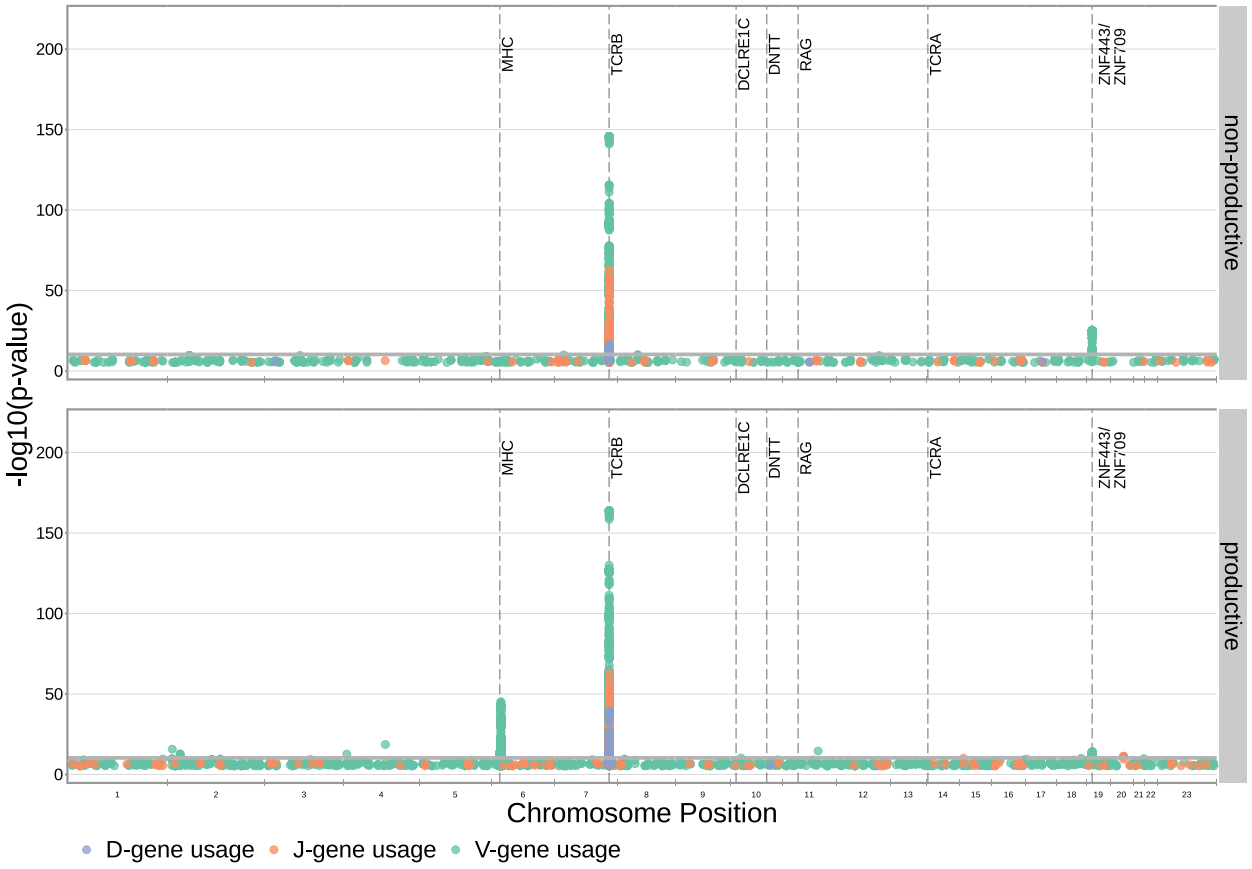

Figure 2.1: Many strong associations are present between V-, D-, and J-gene usage frequency and various SNPs genome-wide for both productive and non-productive TCRs. The most significant SNP associations for the frequency of each of the 60 V-genes, 2 D-genes, and 14 J-genes are located within the *TRB* and MHC loci. Associations are colored by gene-type instead of by gene identity for simplicity. Only SNP associations whose $P < 5 \times 10^{-6}$ are shown here. The gray horizontal line corresponds to a Bonferroni-corrected P-value significance threshold of 5.09×10^{-11}.

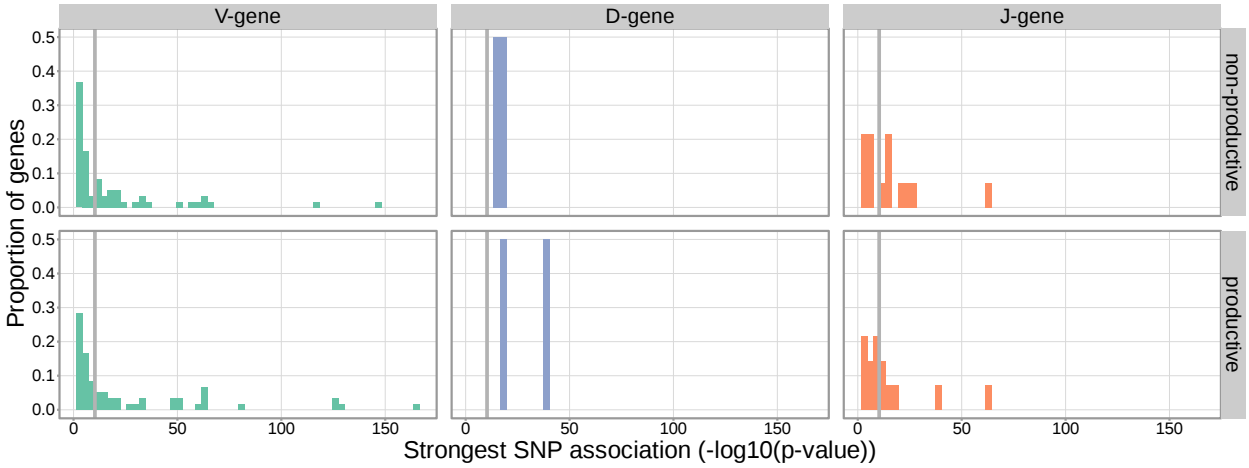

Figure 2.2: Gene-usage frequency of many V-gene, D-gene, and J-gene segments is significantly associated with variation in the *TRB* locus. The P-value of the strongest *TRB* SNP, gene-usage association for each different V-gene, D-gene, and J-gene segment is given on the X-axis. The proportion of gene segments within each gene type is given on the Y-axis. The gray vertical lines correspond to a whole-genome-level Bonferroni-corrected P-value significance threshold of 5.09×10^{-11}.

SNP genotype and gene frequency across all individuals. We fit this "simple model" for each different V-, D-, and J-gene in our paired dataset.

Because of the potential for population-substructure-related effects to inflate associations between each SNP and gene usage frequency, we incorporated ancestry-informative principal components [58] based on the SNP genotypes for a subset of representative subjects as covariates in each model (see Methods for details). Diagnostic statistics show that this bias correction is sufficient.

With these methods, we consider the significance of associations at a Bonferroni-corrected whole-genome P-value significance threshold of 5.09×10^{-11} (see Methods). Using this conservative threshold, we identified 9,152 significant associations between the frequency of various V-, D-, and J-genes and the genotype of SNPs genome-wide (Figure 2.1). Of these significant associations, 7,096 were located within the *TRB* locus for both productive and non-productive TCRs. The *TRB* gene locus encodes the variable V-, D-, and J-gene segments

which are recombined during V(D)J recombination. In our dataset, there are 60 V-genes, 2 D-genes, and 14 J-genes uniquely expressed. As we would expect, we find that the expression of many of these genes is associated with variation in the *TRB* locus (Figure 2.2). For the significantly associated *TRB* locus SNPs, the median association effect magnitude was largest for the expression of TRBD1 (median effect size = -0.038) followed by the expression of TRBD2 (median effect size = 0.035) and the expression of TRBV28 (median effect size = 0.019) all in productive TCRs (Figure A.1). Variation in the *TRB* locus is most significantly associated (smallest P-value) with expression of the gene *TRBV28* within both productive ($P = 1.41 \times 10^{-164}$) and non-productive ($P = 1.94 \times 10^{-146}$) TCR$\beta$ chains. We identified the largest number of significant associations between variation in the *TRB* locus and expression of the gene *TRBV7-3* within productive TCRβ chains (232 significant associations) and the gene *TRBJ1-2* within non-productive TCRβ chains (290 significant associations).

Beyond the *TRB* locus, we also identified 1,242 significant SNP associations within the major histocompatibility complex (MHC) locus. MHC proteins act by presenting self and foreign peptides to TCRs for inspection. Because of this important role in the functionality of T cells, the TCR-MHC interaction is important for thymic selection. We observe the expression of 12.1% of V-genes for productive TCRs to be associated with variation in the MHC locus. For the significantly associated MHC locus SNPs, the median association effect magnitude was largest for the expression of TRBV4-1 (median effect size = -0.004) followed by the expression of TRBV10-3 (median effect size = 0.0033) (Figure A.2). This associated MHC locus variation is located within sequences which code for canonical, peptide-presenting MHC proteins. For example, the eight most significantly associated SNPs were located within the *HLA-DRB1* gene within the MHC locus. These top SNPs were all associated with the expression of the gene *TRBV10-3* within productive TCRs. As expected, the expression of V-genes for non-productive TCRs is not associated with variation in the MHC locus. Likewise, the expression of D- and J-genes for both productive and non-productive TCRs is

not associated with variation in the MHC locus. These results refine and extend associations found in previous work [38, 39].

We observed just one other long-range association region, in addition to the MHC locus, located in proximity to the *ZNF443* and *ZNF709* loci on chromosome 19. Both of these zinc finger proteins contain KRAB-domains and, thus, likely act as transcriptional repressors [59]. In this region, we observe 138 significant SNP associations for the expression of the V-gene *TRBV24-1*. Of these 138 SNP associations, 76 were associations for *TRBV24-1* expression in non-productive TCRs and 62 were associations for *TRBV24-1* expression in productive TCRs. Significant association between variation near the *ZNF443* locus and expression of *TRBV24-1* in productive TCRs was also noted previously [38]. Because the associations observed here are strongest for non-productive TCRs, this chromosome 19 variation likely influences gene usage during TCR generation steps, as opposed to selection. Variation in proximity to the *ZNF443* and *ZNF709* loci may alter the resulting zinc finger proteins and lead to differential transcriptional repression of a site near *TRBV24*. Because the transcription of unrearranged gene segments influences their recombination potential [60], this difference in repression could subsequently change the usage frequency of the *TRBV24* gene.

2.1.3 DCLRE1C locus variation is associated with the extent of trimming

We hypothesized that SNPs across the genome, particularly those within V(D)J-recombination-associated genes, may influence the extent of TCR nucleotide trimming at V(D)J *TRB* gene junctions. It has been previously observed that the extent of trimming varies by V(D)J *TRB* gene choice (Figure A.7) [15, 25, 27, 30]. In other words, two different V-genes (*TRBV19* and *TRBV20-1*, for example) will on average be trimmed to different extents due, in part, to differences in their terminal nucleotide sequences (and the same is true for D- and J-genes). Thus, to quantify the effect of SNPs on the extent of V-, D-, and J-gene trimming during V(D)J recombination, without confounding the extent of trimming with *TRB* gene

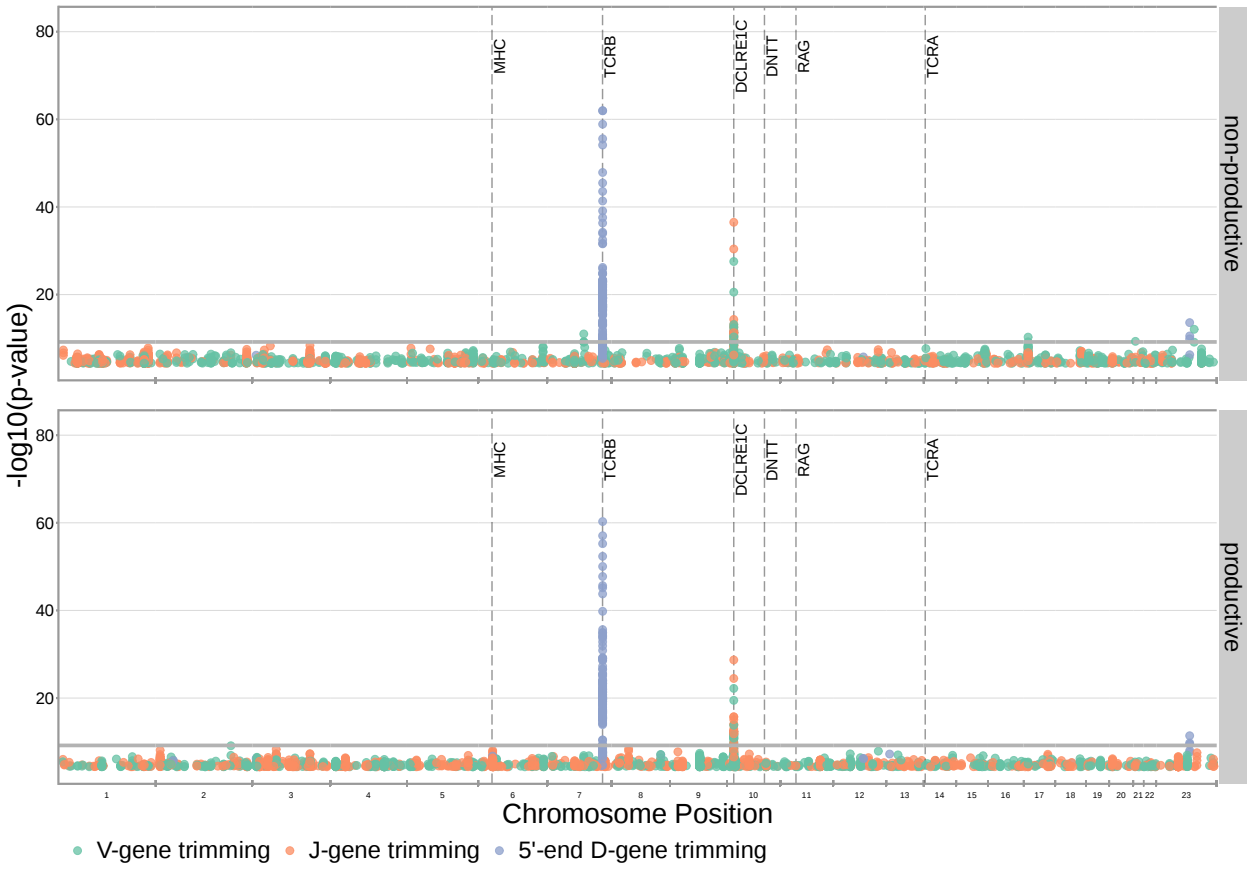

Figure 2.3: SNP associations for all four trimming types reveal the most significant associations to be located within the *TRB* and *DCLRE1C* loci for 5' D-gene trimming and V-gene trimming, respectively, when conditioning out effects mediated by gene choice when calculating the strength of association. Only SNP associations whose $P < 5 \times 10^{-5}$ are shown here. The gray horizontal line corresponds to a Bonferroni-corrected P-value significance threshold of 9.68×10^{-10}.

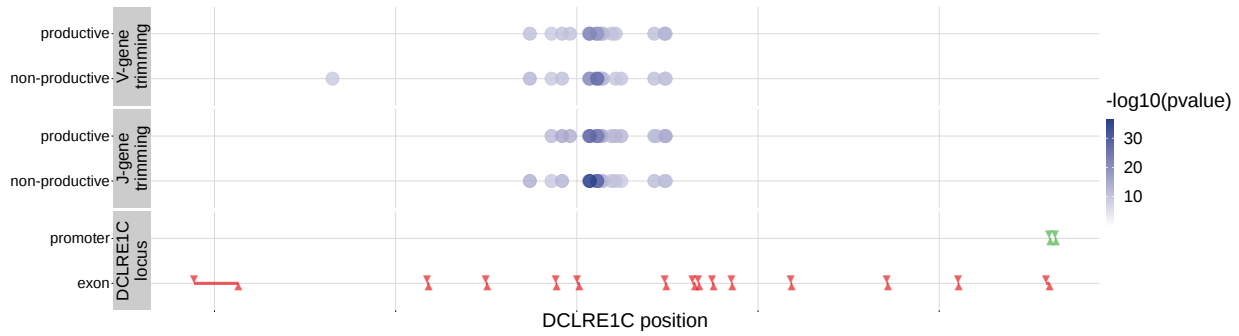

Figure 2.4: Within the *DCLRE1C* locus, 93.8% of these significantly associated SNPs were located within introns. Additionally, many of these significant SNP associations overlapped between trimming types. Downward arrows represent promoter/exon starting positions and upward arrows represent promoter/exon ending positions.

choice, we designed a linear fixed effects model to measure the correlation between each SNP and the number of nucleotide deletions, while conditioning out the effect mediated by gene choice. We fit this "gene-conditioned model" for each of the four trimming types (V-gene trimming, 5' end D-gene trimming, 3' end D-gene trimming, and J-gene trimming) on our paired data set. We performed the analysis, as above, incorporating ancestry-informative principal components in each model (detailed in Methods). Diagnostic statistics show that this correction for population-substructure-related biases is sufficient. Here, we considered the significance of associations at a Bonferroni-corrected whole-genome P-value significance threshold of 9.68×10^{-10} (see Methods).

With these methods, we identified 317 significant SNP associations with the extent of nucleotide trimming for various trimming types (Figure 2.3). We found 66 highly significant associations between V- and J-gene trimming and SNPs within the *DCLRE1C* gene locus for both productive and non-productive TCRs when considered in the whole-genome context. For these significant *DCLRE1C* locus SNP associations, the magnitudes of the effects were greater for non-productive TCRs compared to productive TCRs for both V-gene trimming and J-gene trimming (Figure A.11). The *DCLRE1C* gene encodes the Artemis protein, an endonuclease responsible for cutting the hairpin intermediate prior to nucleotide

trimming and insertion during V(D)J recombination. Many of the SNPs responsible for these 66 significant associations within the *DCLRE1C* locus were shared between trimming and productivity types (Figure 2.4). The most significantly-associated SNP (rs41298872) within this locus had a P-value of 3.18×10^{-37} for J-gene trimming of non-productive TCRs (Figure A.12). This SNP was also significantly-associated with J-gene trimming of productive ($P = 1.99 \times 10^{-29}$) TCRs and V-gene trimming of productive ($P = 6.23 \times 10^{-23}$) and non-productive ($P = 2.81 \times 10^{-21}$) TCRs. We performed a conditional analysis to identify potential independent secondary signals by including this SNP as an additional covariate within the model. This analysis revealed a second, independent SNP signal (rs35441642) within the *DCLRE1C* locus for J-gene trimming of non-productive TCRs. None of the other nucleotide trimming type, productivity status combinations had significant evidence for secondary independent signals.

Our procedure also identified many highly significant associations between 5' end D-gene trimming and SNPs within the *TRB* gene locus, however these appear to result from correlations between SNP genotype and *TRBD2* allele genotype (Figure A.4). If we correct for *TRBD2* allele genotype in our model formulation (see Methods), we no longer observe these associations between SNPs within the *TRB* gene locus and the extent of 5' end D-gene trimming (Figure A.5). *TRBD2* allele genotype could be acting as a confounding variable due to linked local genetic variation which influences nucleotide trimming and/or D-gene assignment ambiguity variation as a function of *TRBD2* allele genotype. To explore the extent of possible D-gene assignment ambiguity variation, we restricted our analysis to TCRs which contain *TRBJ1* genes (and consequently contain *TRBD1* due to topological constraints during V(D)J recombination [19, 61]). With this approach, we also no longer observe associations between SNPs within the *TRB* gene locus and the extent of 5' end D-gene trimming, and additionally, we do observe significant associations between SNPs within the *DCLRE1C* locus and 5' and 3' end D-gene trimming which were not observed in the

original genome-wide analysis (Figure A.6).

Our fixed effects model formulation for these inferences is important: if we don't condition on gene choice then additional, and presumably spurious, associations arise. Indeed, when implementing the "simple model" designed to quantify the association between the four trimming types and genome-wide SNP genotypes, without conditioning out the effect mediated by gene choice, we observe additional associations between SNPs within the MHC locus and V-gene trimming of productive TCRs and between SNPs within the *TRB* locus and V-gene and 3' end D-gene trimming of, again, productive TCRs (Figure A.8). This is perhaps not surprising, as we noted earlier that variations in the MHC and *TRB* loci are associated with gene usage frequencies in productive TCRs (Figure 2.1), and different genes have different trimming distributions (determined in part by the nucleotide sequences at their termini).

Because P-nucleotides can be present at V(D)J junctions in the absence of nucleotide trimming [61], we hypothesized that similar *DCLRE1C* locus variation may also be associated with P-addition. Interestingly, we did not identify any strong associations between SNPs within the *DCLRE1C* locus and the fraction of non-gene-trimmed TCRs containing P-nucleotides when implementing our "gene-conditioned model", despite the known role of the Artemis protein in functioning as an endonuclease responsible for cutting the hairpin intermediate, and thus, potentially creating P-nucleotides during V(D)J recombination (Figure A.9). We observe similar results when quantifying the effect of genome-wide SNPs on the number of V-, D-, and J-gene P-nucleotides per TCR (Figure A.10).

2.1.4 DNTT locus variation is associated with the number of N-insertions

Unlike V-, D-, or J-gene nucleotide trimming length, the number of nucleotide N-insertions between V-D and D-J genes does not vary substantially with V(D)J *TRB* gene choice (Figure A.16) [15]. Thus, to infer the association between SNPs and the number of nucleotide

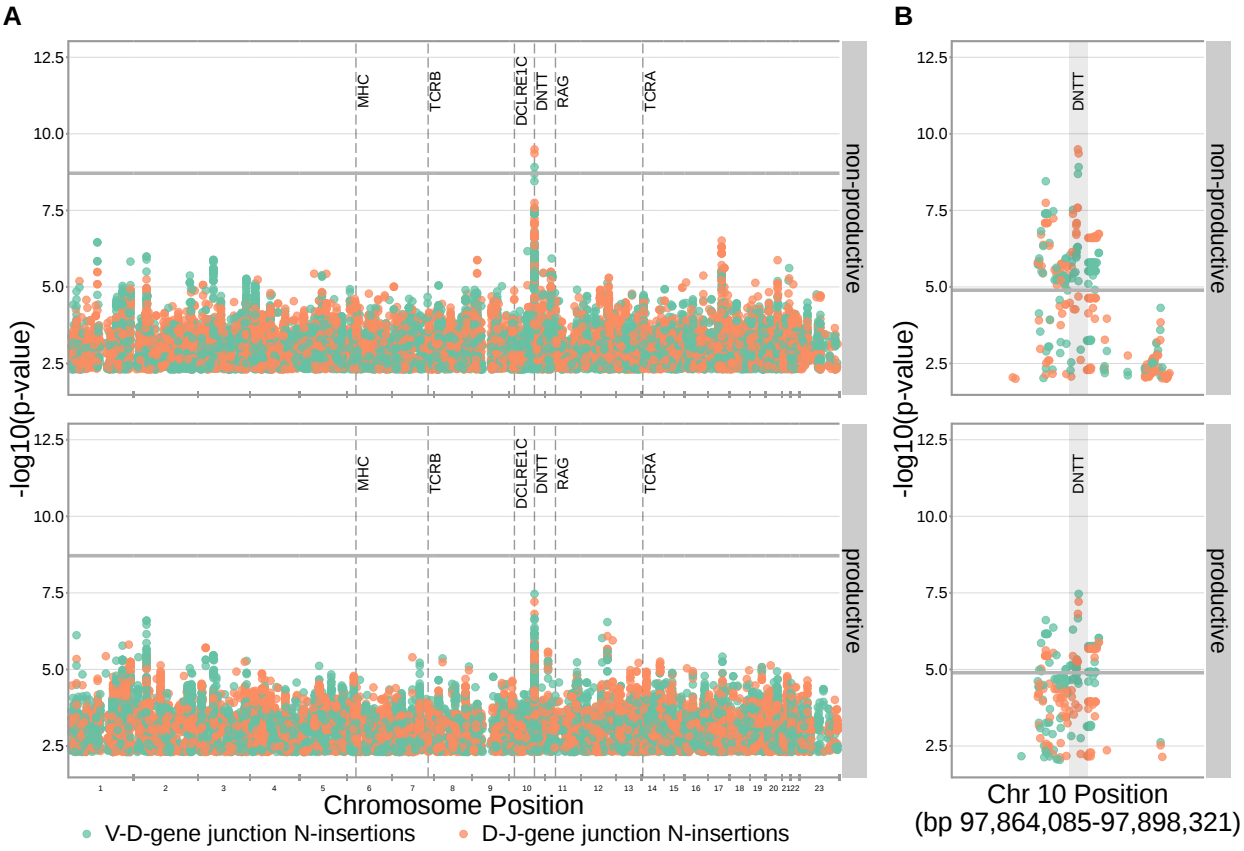

Figure 2.5: SNPs within the *DNTT* locus are associated with the extent of N-insertion. (**A**) There are three associations for SNPs within the *DNTT* locus which are significant when considered in the whole-genome context. The gray horizontal line corresponds to a whole-genome Bonferroni-corrected P-value significance threshold of 1.94×10^{-9}. (**B**) Using a *DNTT* gene-level significance threshold, many more SNPs within the extended *DNTT* locus have significant associations for both N-insertion types. Here, the gray horizontal line corresponds to a gene-level Bonferroni-corrected P-value significance threshold of 1.28×10^{-5} (calculated using gene-level Bonferroni correction for the 977 SNPs within 200kb of the *DNTT* locus, see Methods). For both (A) and (B), only SNP associations whose $P < 5 \times 10^{-3}$ are shown.

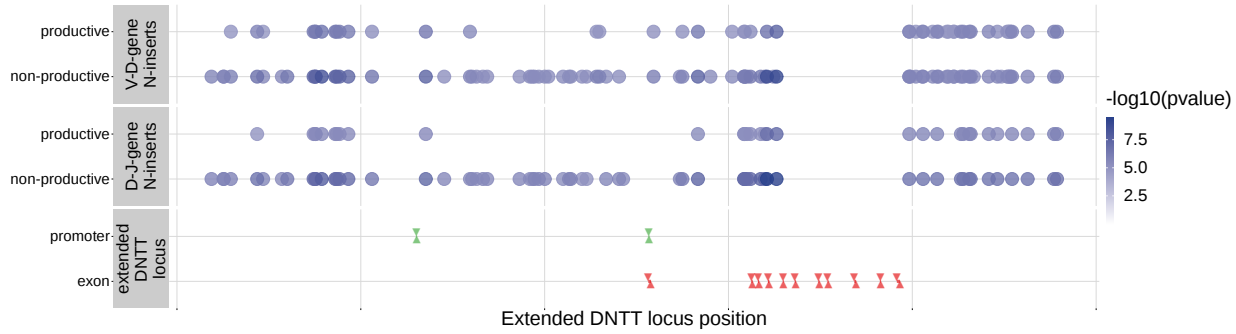

Figure 2.6: Within the *DNTT* locus, many of the significant SNP associations overlapped between N-insertion types when using *DNTT* gene-level Bonferroni-corrected P-value significance threshold of 1.28×10^{-5}. Downward arrows represent promoter/exon starting positions and upward arrows represent promoter/exon ending positions.

N-insertions, we implemented a "simple model", without conditioning out any effect mediated by gene choice. Again, because of the potential for population-substructure-related effects to inflate associations between each SNP and the number of N-insertions, we incorporated ancestry-informative principal components as covariates in each model (detailed in Methods). Diagnostic statistics show that this bias correction is sufficient.

With these methods, we identified three associations between SNPs and the number of nucleotide N-insertions using a Bonferroni-corrected whole-genome P-value significance threshold of 1.94×10^{-9} (see Methods) (Figure 2.5). Two SNPs within the *DNTT* gene locus (rs2273892 and rs12569756) were responsible for these associations. The *DNTT* gene encodes the terminal deoxynucleotidyl transferase (TdT) protein which is a specialized DNA polymerase responsible for adding non-templated (N) nucleotides to coding junctions during V(D)J recombination. When we restrict our analysis to TCRs which contain *TRBJ1* genes and consequently eliminate potential D-gene assignment ambiguity, we continue to observe these *DNTT* associations (Figure A.17).

Since the TdT protein has an important mechanistic role in the N-insertion process and because we already identified SNPs within the *DNTT* locus to be weakly associated with the number of N-insertions at V(D)J gene junctions, we wanted to explore the locus further.

Restricting the analysis to the extended *DNTT* locus reduced the multiple testing burden such that 232 significant associations emerged (Figure 2.5). For these significant *DNTT* locus SNP associations, the magnitudes of the effects were greater for non-productive TCRs compared to productive TCRs for both V-D-gene junction N-insertion and D-J-gene junction N-insertion (Figure A.18). Many of the SNPs responsible for these 232 significant associations within the extended *DNTT* locus were shared between insertion and productivity types (Figure 2.6). While most of these associations are likely the result of a single independent signal for each insertion and productivity type, we performed a conditional analysis to identify potential independent secondary signals. To do so, we included the most significant SNP within the *DNTT* locus for each insertion and productivity type as a covariate in the model. With this approach, we identified rs2273892 as the primary independent signal for D-J N-insertion of non-productive TCRs and rs12569756 as the primary independent signal for D-J N-insertion of productive TCRs and V-D N-insertion of productive and non-productive TCRs. However, these two SNPs are tightly linked and, thus, likely both represent the same, primary independent signal. This analysis did not reveal any significant evidence for secondary independent signals.

We found that correcting for population-substructure-related effects was especially important in our primary genome-wide analysis, which led us to discover differences in the extent of N-insertion by ancestry-informative PCA cluster. Indeed, if we don't incorporate correction terms for population-substructure-related biases in our model formulation, we observe many strongly significant associations, particularly within the *DNTT* locus. This hinted at important PCA-cluster level effects. When we look closely at the average number of N-insertions (combining the number of V-D and D-J N-insertions) across TCR repertoires by PCA cluster, we note that subjects from the "Asian"-associated PCA cluster have significantly fewer total N-insertions for productive ($P = 0.006$ without Bonferroni correction) and non-productive ($P = 0.014$ without Bonferroni correction) TCRs when compared to the

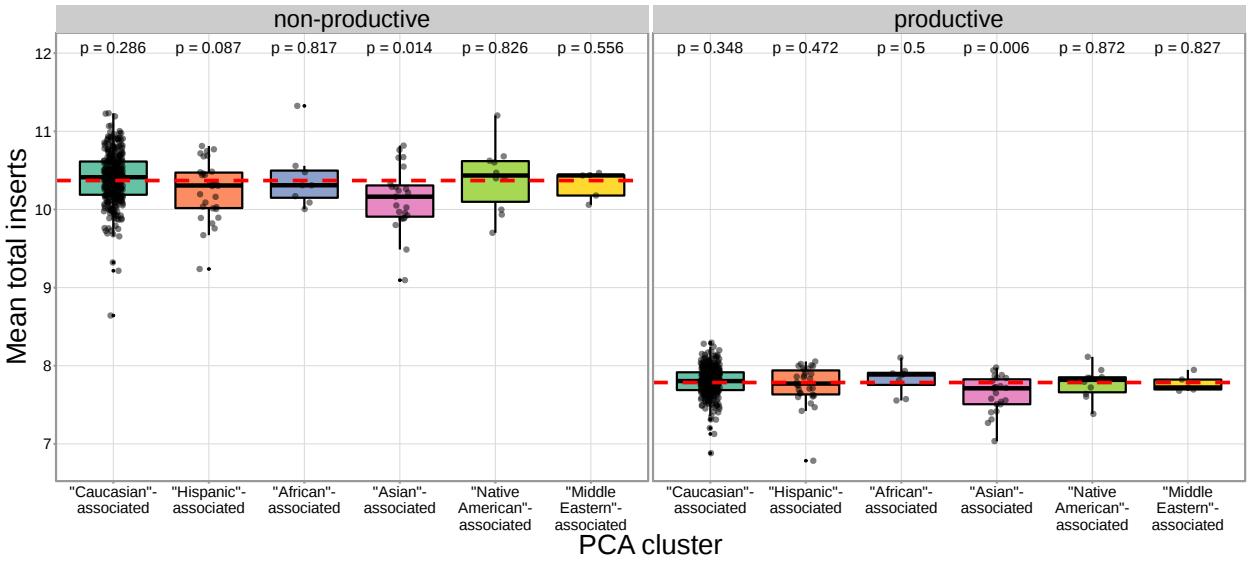

Figure 2.7: The TCR repertoires for subjects in the "Asian"-associated PCA-cluster contain fewer N-insertions for productive TCRs when compared to the population mean computed across all 666 subjects (dashed, red horizontal line). The P-values from a one-sample t-test (without Bonferroni multiple testing correction) for each PCA cluster compared to the population mean are reported at the top of the plot.

population mean (using a one-sample t-test) (Figure 2.7). The total N-insertions for productive TCRs within the "Asian"-associated PCA cluster remain significantly different from the population mean after Bonferroni multiple testing correction (corrected $P = 0.036$). Furthermore, the "Asian"- and "Hispanic"-associated PCA clusters had significantly higher mean SNP allele frequencies for SNPs within the extended *DNTT* region that were associated with fewer N-insertions when compared to the mean population allele frequency ($P = 7.32 \times 10^{-20}$ for the "Asian"-associated PCA cluster and $P = 1.17 \times 10^{-5}$ for the "Hispanic"-associated PCA cluster using a one-sample t-test with Bonferroni multiple testing correction) (Figure 2.8).

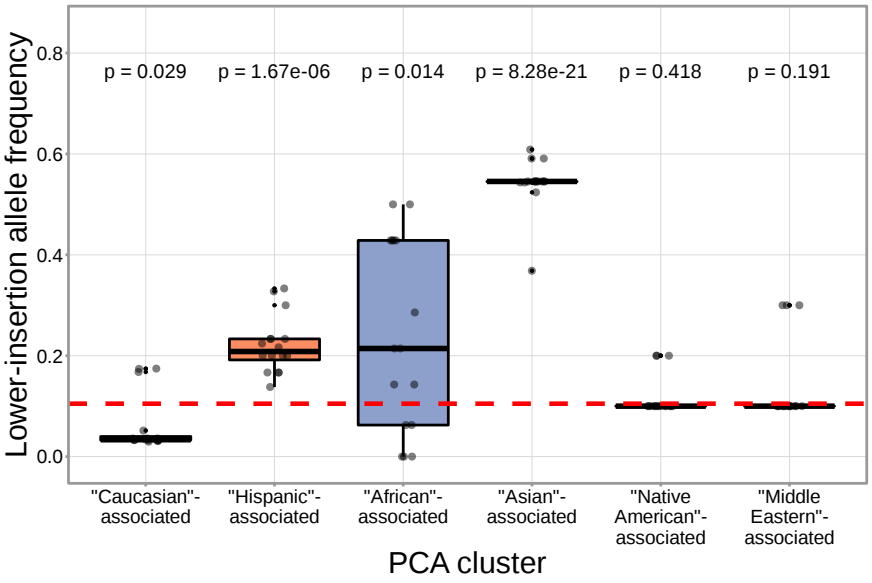

Figure 2.8: SNPs within the *DNTT* region that are associated with fewer N-insertions have a higher mean allele frequency within the "Asian"-associated PCA-cluster when compared to the population mean allele frequency computed across the 398 discovery cohort subjects (dashed, red horizontal line). The P-values from a one-sample t-test (without Bonferroni multiple testing correction) for each PCA cluster compared to the population mean are reported at the top of the plot. The population mean is dominated by subjects in the "Caucasian"-associated PCA cluster (Figure A.22).

2.1.5 Validation Analysis

To validate our results, we worked with paired ancestry-informative marker (AIM) SNP array and TCRα- and TCRβ-immunosequencing data representing 94 individuals and 2 SNPs (which overlap with the discovery dataset) from an independent validation cohort (Table 2.3 and see Methods). In contrast to the discovery cohort, this cohort contains different demographics, shallower RNA-seq based TCR-sequencing, and a sparser set of SNPs. However, TCR-sequencing for both TCRα and TCRβ chains is available.

We were able to validate a discovery-cohort significantly associated *DCLRE1C* SNP within this validation cohort. While none of the independent *DCLRE1C* SNPs from the discovery-cohort analysis overlapped with the validation cohort SNP set, a single, non-

Table 2.3: Validation cohort demographics.

		Count
Sex	Female	58
	Male	36
Age (in years)	< 10	26
	11-20	15
	21-30	13
	31-40	12
	41-50	11
	51-60	9
	> 60	8
Self-reported ethnicity	Hispanic or Latino	94
CMV serostatus	Positive	37
	Negative	57
Total		94

synonymous SNP (rs12768894, c.728A>G) within the *DCLRE1C* locus was present in both SNP sets. This SNP was one of the significant associations we observed for V-gene trimming (productive $P = 2.16 \times 10^{-14}$; non-productive $P = 7.21 \times 10^{-14}$) and J-gene trimming (productive $P = 1.23 \times 10^{-11}$; non-productive $P = 6.62 \times 10^{-12}$) of TCRβ chains in the genome-wide discovery cohort analysis (Figure A.13). Using the same methods, we identified significant associations between this SNP and J-gene trimming of productive TCRα and TCRβ chains and V-gene trimming of both productive and non-productive TCRα and TCRβ chains within the validation cohort (Table 2.4, Figure A.14, and Figure A.15). Associations between rs12768894 and both types of D-gene trimming of TCRβ chains were not significant for either cohort.

We were unable to validate the most significantly associated *DNTT* SNPs due to lack of overlap between the SNP sets for the discovery and validation cohorts; a discovery-cohort weakly associated SNP (rs3762093) failed to reach statistical significance for all N-insertion types, but had the same direction of effect in the validation cohort as follows. Within the discovery cohort, rs3762093 genotype was weakly associated with the number of V-D

N-insertions (productive $P = 1.37 \times 10^{-6}$; non-productive $P = 1.50 \times 10^{-7}$) and D-J N-insertions (productive $P = 9.43 \times 10^{-6}$; non-productive $P = 1.94 \times 10^{-7}$) within TCRβ chains (Figure A.19). Within the validation cohort, this SNP was significantly associated with the number of V-J N-insertions within productive TCRα chains (Table 2.4 and Figure A.21). However, this SNP was not significantly associated with the number of V-D or D-J N-insertions within productive or non-productive TCRβ chains or the number of V-J N-insertions within non-productive TCRα chains within the validation cohort (Table 2.4, Figure A.20, and Figure A.21). Despite the lack of significance, we noted that the model coefficients for rs3762093 genotype were in the same direction (i.e., the minor allele was associated with fewer N-insertions) for all N-insertion and productivity types within TCRβ chains for both cohorts. Further, while TCRα chain sequencing was not available for the discovery cohort, we observed stronger associations between rs3762093 genotype and the extent of N-insertion for both productivity types within TCRα chains compared to TCRβ chains within the validation cohort. Perhaps with a larger validation cohort, significant associations would be present for all N-insertion types.

2.2 Discussion

V(D)J recombination is a complex stochastic process that enables the generation of diverse TCR repertoires. Our results show that genetic variation in various V(D)J recombination genes has a key role in shaping the TCR repertoire through biasing V(D)J gene choice, nucleotide trimming, and N-insertion in a broad population sample. While we recognize that there may be a complicated entanglement between allelic variation and local *cis*-acting effects, we were primarily interested in identifying strong, *trans*-acting associations. By leveraging the unique pairing of TCRβ chain immunosequencing and genome-wide genotype data, we have (1) confirmed and extended previous studies on the genetic determinants of

Table 2.4: We inferred the associations between SNP genotype and TCR repertoire features for two SNPs overlapping between discovery-cohort and validation-cohort SNP sets. We considered the significance of the validation cohort associations at a Bonferroni-corrected SNP-level P-value significance threshold of 0.0042 for trimming and 0.0083 for N-insertion (see Methods). Validation cohort P-values are one-tailed. * discovery-cohort associations were only significant when considered at the *DNTT*-gene level significance threshold, not at the whole-genome significance threshold.

SNP	TCR chain	Repertoire feature	Sequence type	Discovery cohort signif. association	Validation cohort signif. association
rs12768894	TCRβ	V-gene trimming	Productive	Yes (2.16×10^{-14})	Yes (7.17×10^{-7})
			Non-productive	Yes (7.21×10^{-14})	Yes (8.75×10^{-6})
		J-gene trimming	Productive	Yes (1.23×10^{-11})	Yes (5.16×10^{-10})
			Non-productive	Yes (6.62×10^{-12})	No (4.18×10^{-2})
	TCRα	V-gene trimming	Productive	N/A	Yes (2.59×10^{-5})
			Non-productive	N/A	Yes (2.68×10^{-7})
		J-gene trimming	Productive	N/A	Yes (6.29×10^{-12})
			Non-productive	N/A	No (9.99×10^{-3})
rs3762093	TCRβ	V-D N-insertion	Productive	Yes* (1.37×10^{-6})	No (0.153)
			Non-productive	Yes* (1.50×10^{-7})	No (0.059)
		D-J N-insertion	Productive	Yes* (9.43×10^{-6})	No (0.137)
			Non-productive	Yes* (1.94×10^{-7})	No (0.067)
	TCRα	V-J N-insertion	Productive	N/A	Yes (0.006)
			Non-productive	N/A	No (0.031)

TCR V-gene usage, (2) discovered associations between common genetic variants within the *DCLRE1C* and *DNTT* loci and V(D)J junctional trimming and N-insertions, respectively, (3) developed a method for quantifying the extent of the associations between genetic variations and junctional features, directly, without confounding gene choice effects, and (4) revealed differences in the extent of N-insertion by ancestry-informative PCA cluster.

We note an abundance of associations between variation in the *TRB* locus and V(D)J gene usage biases for both productive and non-productive TCRs. Although previous reports have revealed similar patterns of association for productive TCRs [38, 39], our results refine and extend this result by quantifying the extent of *TRB* locus variation on V(D)J gene usage for non-productive TCRs. This highlights that locus variation is associated with

TCR generation-related gene usage biases, in addition to potential thymic selection biases for productive TCRs. These TCR generation-related gene usage biases likely reflect local gene regulation and/or recombination efficiency effects. For example, one of the SNPs most significantly associated with *TRBV28* expression (rs17213) is located within the recombination signal sequence at the 3'-end of the gene and, thus, could be involved directly in changing the recombination efficiency of *TRBV28*. Thus, different expression levels of various genes could be promoted by variation within non-coding regions such as promoters, 5'UTRs and leader sequences, introns, or recombination signal sequences. Polymorphisms within these regions have been suggested to influence V(D)J gene expression levels within B-cell receptor repertoires [62]. We also observed that variation in the MHC locus is associated with V-gene usage biases for productive TCRs, but not non-productive TCRs. These MHC locus associations are likely only observed for V-gene usage since the V-gene locus, exclusively, encodes the TCR regions (complementarity-determining regions 1 and 2) which directly contact MHC during peptide presentation [61]. While significant associations between MHC locus variation and V-gene usage have been identified previously [38, 39], the specific MHC locus variants and V-genes responsible for the most significant of these associations differed between the two studies and from those reported here. This variation is likely the result of population composition and/or exposure history differences between the various study cohorts. Despite their differences, both previous studies have suggested that the thymic selection of certain V-genes may be biased by germline-encoded TCR-MHC compatibilities in an MHC dependent manner [38, 39]. Because of our observed distinction between associations present between MHC variation and V-gene usage in productive versus non-productive TCRs, our work supports this hypothesis.

We have identified, for the first time, specific genetic variants which are associated with modifying the extent of N-insertion and nucleotide trimming. While many previous studies have reported evidence of genetic influences on overall gene usage [40, 41, 53–56] and reper-

toire similarity in response to acute infection [41, 54], there have been few explorations into how heritable factors may bias TCR junctional features beyond reports of genetic similarity implying overall TCR repertoire similarity [40, 51]. Here, we noted that variation in the gene encoding the Artemis protein (*DCLRE1C*) is associated with the extent of V- and J-gene nucleotide trimming for both productive and non-productive TCRs. These associations are strongest for non-productive TCRs suggesting a TCR generation-related repertoire bias. It is well established that the Artemis protein, in complex with DNA-PKcs, functions as an endonuclease responsible for cutting the hairpin intermediate, and thus, potentially creating P-nucleotides prior to nucleotide trimming during V(D)J recombination [24, 28, 29, 31]. The direct involvement of Artemis in the nucleotide trimming mechanism, however, has yet to be confirmed. It has been shown that the Artemis protein possesses single-strand-specific 5' to 3' exonuclease activity [29, 63] and, thus, may be properly positioned to trim nucleotides. A non-synonymous SNP within *DCLRE1C* (rs12768894, c.728A>G) was one of the significant associations we observed for V- and J-gene nucleotide trimming in both the primary cohort and the independent validation cohort. Perhaps this mutation, or other linked non-synonymous *DCLRE1C* variation that was not studied here, is directly involved in the trimming changes we observe. We did not observe strong associations between variation in the *DCLRE1C* locus and the number of P-nucleotides or the fraction of non-gene-trimmed TCRs containing P-nucleotides, despite the established mutually exclusive relationship between P-addition and nucleotide trimming [26, 47, 61]. However, the absence of P-nucleotide associations at the *DCLRE1C* locus could be the result of restricting the analyses to the non-gene-trimmed repertoire subset. Perhaps with a larger dataset these associations would be present.

Further, we have identified associations between variation in the gene encoding the TdT protein (*DNTT*) and the number of N-insertions for both productive and non-productive TCRs. Because of the established, direct involvement of the TdT protein in the N-insertion

mechanism, these *DNTT* locus variations could be influencing the function of the TdT protein. These significant associations were slightly stronger for non-productive TCRs perhaps suggesting that thymic selection may limit the mechanistic effects of locus variation. Interestingly, we noted that the extent of N-insertion varies by ancestry-informative PCA cluster. Specifically, we found that the "Asian"-associated PCA cluster had significantly fewer N-insertions for productive TCRs when compared to the population mean which is dominated by the "Caucasian"-associated PCA cluster. This finding is, perhaps, related to the influence of broad heritable factors biasing the extent of N-insertions.

The significant SNPs associated with changing the extent of nucleotide trimming and N-insertion identified here could be expression quantitative trait loci (eQTLs), however, experimental work will be required to determine whether these SNPs modify the expression of *DCLRE1C*and *DNTT*, respectively. More work is also required to elucidate the mechanistic relationship between *DCLRE1C* locus variation and nucleotide trimming changes. After characterizing these relationships, future work can focus on identifying correlations between TCR repertoires and host immune exposures while accounting for genetically determined repertoire biases identified here. These directions would allow us to continue disentangling the genetic and environmental determinants governing TCR repertoire diversity.

There are several key limitations of our approach which are intrinsic to the data used in this study. First, the lack of overlap between SNP sets for the discovery and validation cohorts limited our ability to directly validate our strongest inferences. Next, it is possible that the SNP array data used here does not capture all potential causal variation. As such, a significantly associated SNP present in our SNP array data could simply be in linkage disequilibrium with a causal SNP which was either poorly imputed or not tested here. Previous work has suggested that polymorphisms within the immunoglobulin V-gene region are not completely captured by existing SNP array technology, and have been underrepresented in previous genome-wide association studies [64]. SNP coverage of the *TRB* locus is

thought to be even sparser [65], and thus, much of the actual *TRB* variation present within our data cohort is likely not captured by the SNP dataset used here (which contains 7,304 SNPs within the TRB locus, hg19:chr7:141950000-142550000). Lastly, we have used the recombination statistics from non-productive rearrangements here as a means of studying the V(D)J recombination generation process in the absence of selection, however, we acknowledge that the repertoire of non-productive rearrangements may be an imperfect proxy for a pre-selection TCR repertoire. Since each non-productive rearrangement is sequenced due to the presence in the same T cell of a successful rearrangement that survived selection, it is possible that within-cell correlation between rearrangement events could imprint selection effects onto the non-productive repertoire. However, we are not aware of any evidence for a mechanism in which productive and non-productive recombination events at the *TRB* locus are significantly correlated. As such, we are assuming that the productive and non-productive recombination events are independent, and thus, the recombination statistics from the repertoire of non-productive rearrangements should reflect that of a pre-selection repertoire as is common in the literature [15, 19, 40, 41, 53].

Another key constraint is the challenge of inferring the V(D)J rearrangements from the final nucleotide sequences due to the poor characterization of the *TRA* and *TRB* loci. The *TRA* and *TRB* regions have been historically difficult to reliably map using short read sequencing due to their repetitive and complex nature. While recent work has identified many new *TRBV* alleles, many more undocumented *TRBV* alleles likely remain to be discovered [65]. As such, the incomplete characterization of the *TRB* locus limited our ability to infer the correct V(D)J -gene allele for each final nucleotide sequence. Further, the TCR sequencing technology used here leverages relatively short-read sequencing which captures only a portion of the V-gene present in each sequence. Because many *TRBV* alleles are identical to other *TRBV* alleles for much of the V-gene region present in these sequences, it can be difficult to unambiguously assign V-gene usage to the final nucleotide sequences. D-

gene usage assignment is also challenging due to the short length of the *TRBD* alleles (12-16 nucleotides before nucleotide trimming and N-insertion). We have found that controlling for D-gene assignment ambiguity in the nucleotide trimming and N-insertion analyses results in similar significant associations within the *DNTT* and *DCLRE1C* loci. Although we cannot rule out some effect of incorrect V(D)J -gene assignment bias for *trans* associations resulting from the signal being "masked" by stronger *TRB* locus signals, these biases seem to be mostly restricted to *cis* associations.

In summary, we have found that the usage of *TRB* genes is associated with variation in MHC and *TRB* loci, the number of N-insertions is associated with *DNTT* variation, and the extent of nucleotide trimming is associated with *DCLRE1C* variation. Our results clearly demonstrate how variation in V(D)J recombination-related genes can bias TCR repertoire combinatorial and junctional diversity. In the case of B cells, genetically determined V(D)J gene usage biases within B-cell receptor repertoires have been linked to functional consequences for the overall immune response to specific antigens and, thus, an increased susceptibility to certain diseases [62]. As such, the genetic TCR repertoire biases identified here lay the groundwork for further exploration into the diversity of immune responses and disease susceptibilities between individuals. Such studies will enhance our understanding of how an individual's diverse TCR repertoire can support a unique, robust immune response to disease and vaccination. Our findings also provide a step towards the ability to understand and predict an individual's TCR repertoire composition which will be critical for the future development of personalized therapeutic interventions and rational vaccine design.

2.3 Methods and Materials

2.3.1 Data details

Discovery cohort dataset: TCRβ repertoire sequence data for 666 healthy bone marrow donor subjects was downloaded from the Adaptive Biotechnologies website using the link provided in the original publication [50]. For both the discovery and validation cohorts, V, D, and J genes were assigned by comparing the TCRβ-chain (and TCRα-chain for the validation cohort) nucleotide sequences to the human IMGT/GENE-DB *TRB* (or *TRA*) allele sequences [66]. To infer the extent of nucleotide trimming, N-insertion, and P-addition for each TCRβ-chain (and TCRα-chain) nucleotide sequence, the most parsimonious V(D)J recombination scenario was assigned to each sequence using the TCRdist pipeline [67]. The V(D)J recombination scenario requiring the fewest N-insertions was defined as the most parsimonious scenario.

SNP array data corresponding to 398 of these subjects was downloaded from The database of Genotypes and Phenotypes (accession number: phs001918). Details of the SNP array dataset, genotype imputation, and quality control have been described previously [57].

Validation cohort dataset: Peripheral blood mononuclear cell (PBMC) samples were collected from 150 healthy subjects recruited at the Health Center Sócrates Flores Vivas (HCSFV) in Managua, Nicaragua [68]. Healthy participants were recruited as contacts of influenza infected index patients and blood samples were collected at both the initial visit and a 30 day follow-up visit. Participants provided written informed consent and parental permission was obtained from parents or legal guardians of children, in addition to verbal assent from children aged six years and older. This study was approved by the Institutional Review Boards at the University of Michigan (HUM 00091392) and the Centro Nacional de Diagnóstico y Referencia (Ministry of Health, Nicaragua; CIRE 06/07/10-025).

With these samples, PBMCs were stained with CD3-PerCP eFluor710 (Thermo Cat. 46-0037-42), CD4-BV650 (BD Biosciences Cat. 563875), CD8-APC Fire750 (Biolegend Cat. 344746), and gdCy7 (Biolegend Cat. 331222). Briefly, after thawing from cryopreservation and plating in a 96-well round bottom plate, cells were spun down and resuspended in 50 μL of human Fc block (BD Biosciences Cat. 564220) in Dulbecco's phosphate-buffered saline (DPBS) at 5 μL per test (1 test = 1.0×10^6 cells) and incubated for 10 minutes at room temperature. Afterwards, 50 μL of a 2X Live/Dead Aqua (Tonbo Cat. 13-0870-T100, 1 μL per test, 1 test = 1.0×10^6 cells) and pre-titrated surface antibody cocktail in DPBS were added to each well and cells were incubated for 30 minutes on ice and in the dark. Cells were washed, resuspended in sort buffer and bulk sorted into polystyrene tubes. Afterwards, samples were spun down, pellets were resuspended in 350 μL of RNA lysis buffer, and stored at -80 C in labeled epitubes.

From here, DNA was extracted from 200 μL of neutrophil pellets using the Qiagen QIAamp DNA Mini Kit (Cat. 51306). Bulk repertoires for sorted CD4 and CD8 T cells were generated in accordance with the protocol developed by [69], and sequencing was performed on the NovaSeq by the Hartwell Center at St. Jude. Raw cDNA sequencing data were processed with the MIGEC software package [70] to define error-corrected *TRA* and *TRB* transcript sequences, which were then analyzed as described above for the discovery cohort data [50].

Genotypes for SNPs of interest corresponding to 94 of these subjects were pulled from Infinium Global Screening Array-24 v3.0 BeadChip results, which measures 654,027 SNP makers including multi-ethnic genome-wide content, curated clinical research variants, and quality control markers. High quality DNA was extracted using the Qiagen QIAamp DNA Mini Kit (Cat. 51306), and submitted to the St. Jude Hartwell Center for preparation and processing. Two SNPs, rs72640001 and rs72772435, were not included on this chip and were determined using Thermo Fisher TaqMan SNP Genotyping Assays (Cat. 4351379, Assay ID

C__99271581_10 and C__99587751_10, respectively) and TaqMan Genotyping Master Mix (Cat. 4371353) according to the kit manual.

2.3.2 Data preparation

With these paired SNP array and TCR-immunosequencing for both the discovery and validation cohorts, we aimed to identify significant associations between these SNPs and various TCR repertoire features. Because we would expect a difference in these phenotypes depending on whether a TCR sequence is classified as productive or non-productive, we split the data based on this TCR productivity status and computed associations separately for the two groups.

We also subset the SNP data further based on several quality control metrics. We filtered the SNP array data to use only SNPs with a minor allele frequency above 0.05 in our analyses which excluded SNPs for which all subjects had the same genotype. For the discovery cohort, this filtering procedure and previous quality control [57] left 6,456,824 SNPs (of the original 35 million SNPs) remaining for our analyses. Only 2 SNPs from the validation cohort overlapped with this discovery cohort SNP set. For each of these discovery and validation cohort SNPs, when fitting each association model, we excluded observations which contained a missing SNP genotype. Next, for the TCR repertoire data, we excluded repertoires which contained a relatively small number of TCRs ($\log_{10}$(TCR count) < 4.25 for productive TCRs and $\log_{10}$(TCR count) < 3.5 for non-productive TCRs) from the analyses. Also, when fitting models for gene usage (i.e. V-gene usage, D-gene usage, and J-gene usage) we have restricted our analyses to observations which contain non-orphan genes. Lastly, for TCRβ-chains, if a D-gene is trimmed so much that the D-gene is unidentifiable, the inference pipeline used to infer *TRB* genes for each sequenced TCR does not report a D-gene. Instead, this D-gene (if it is indeed present) is reported as a V-J N-insertion. Because of this, we excluded these observations when fitting models for TCR features involving the D-gene (i.e.

D-gene usage, both V-D and D-J junction N-insertions, D-gene P-additions, and D-gene nucleotide trimming).

2.3.3 Notation

The discovery dataset contains observations for a total of $I = 398$ subjects and the validation dataset contains observations for a total of $I = 94$ subjects. Within each cohort, for subject $i \in \{1, \ldots, I\}$, we observe a total of N_i TCRs which, here, represents the number of TCRs which compose each subject's TCR repertoire. Thus, for each TCR $k \in \{1, \ldots, N_i\}$, we measure a TCR feature of interest, y_{ik}, such as the number of V-D N-insertions, the extent of V-trimming, etc. We also have SNP genotype data for a total of J SNPs such that for each SNP $j \in \{1, \ldots, J\}$ and subject $i \in \{1, \ldots, I\}$, we measure the number of minor alleles in the genotype, $x_{ij} \in \{0, 1, 2\}$.

2.3.4 Quantifying associations between SNPs and TCR features using the "simple model"

We first describe what we call the "simple model". We will describe more complex models, as well as each model with added correction for population-substructure-related effects, in the sections following.

We began by calculating the average occurrence of the TCR feature of interest, $\bar{y}_i$, within the repertoire of each subject, i. By condensing the data in this way, for each subject $i \in \{1, \ldots, I\}$, we are left with $N_i = 1$ observations. For example, for the discovery cohort, we can fit the model across $\sum_{i=1}^{I} N_i = 398$ observations. Using this condensed dataset, for each SNP, TCR feature, and productivity status, we can fit the model:

$$\bar{y}_i = x_{ij} \cdot \beta_{1j} + \beta_0 + \epsilon_{ij} \tag{2.1}$$

where β_{1j} is the allele effect for SNP j on the TCR feature of interest $\bar{y}_i$, β_0 is the intercept,

and ϵ_{ij} is the random error for subject i and SNP j such that $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

To estimate each regression coefficient, we solved the least squares problem:

$$(\hat{\beta}_0, \hat{\beta}_{1j}) = \text{argmin}_{\beta_0, \beta_{1j}} \sum_{i=1}^{n} \left(\bar{y}_i - (x_{ij} \cdot \beta_{1j} + \beta_0)\right)^2 \tag{2.2}$$

using the function `lm` in R. With each estimate of the j-th SNP effect on the TCR feature of interest, $\hat{\beta}_{1j}$, generated by fitting the least squares problem (2.2), we quantified the association strength between each SNP and the TCR feature of interest by testing whether $\hat{\beta}_{1j} = 0$. To do this, we calculate the test statistic:

$$T_j = \frac{\hat{\beta}_{1j}}{\text{se}(\hat{\beta}_{1j})} \tag{2.3}$$

and compare T_j to a $N(0, 1)$ distribution to obtain each P-value.

2.3.5 Quantifying associations between SNPs and TCR features, conditioned on TRB gene type, using the "gene-conditioned model"

We noted that the amount of certain TCR features (such as the extent of all types of nucleotide trimming) vary by V(D)J *TRB* gene choice. Thus, we can condition on this gene choice to quantify the direct association between each SNP and the amount of each TCR feature, without confounding gene choice effects. In this way, we condition on each gene type $t \in \{\text{V-gene, J-gene, D-gene}\}$ corresponding to the TCR feature of interest (i.e. $t =$ V-gene for V-gene trimming, $t =$ J-gene for J-gene trimming, etc.). We will refer to the following model as the "gene-conditioned model" in the main text. Many similarities exist between the "simple model" described in the previous section and this "gene-conditioned model." Thus, we will focus on the differences between the two models here. We will describe both models with added correction for population-substructure-related effects, in the sections following.

As in the previous section, we, again, want to reduce the number of data observations.

For each subject $i \in \{1, \ldots, I\}$, we can calculate the average amount of each TCR feature $\bar{y}_{im}$ by each candidate *TRB* gene allele group m for the given gene type t such that $m \in \{1, \ldots, M_t\}$. In calculating the average amount of each TCR feature across TCRs with the same candidate *TRB* gene allele, we combined *TRB* gene alleles which had identical CDR3 sequences and were of the same candidate *TRB* gene into *TRB* gene allele groups. As such, the number of observations per subject N_i in this condensed dataset will equal M_t and, thus, we will need to fit each model across $\sum_{i=1}^{I} M_t$ observations. In our data, for TCRβ chains, we observe 141 possible *TRB* V-gene allele groups, 16 J-gene allele groups, and 3 D-gene allele groups. Thus, using the extent of nucleotide trimming as an example TCR feature within the discovery cohort, with this condensed formulation, for each SNP and productivity status, we have $\sim 56,000$ observations for V-gene trimming, $\sim 6,000$ observations for J-gene trimming, and $\sim 1,200$ observations for both types of D-gene trimming.

Using this condensed dataset, for each SNP, TCR feature, and productivity status, we fit the following "gene-conditioned model":

$$\bar{y}_{im} = x_{ij} \cdot \beta_{1j} + \beta_0 + \gamma_{jm} + \epsilon_{ijm} \tag{2.4}$$

where γ_{jm} represents the gene-effect on the amount of the TCR feature of interest for SNP j and gene-allele-group m, and ϵ_{ijm} is the random error for subject i, SNP j, and gene-allele-group m such that $\epsilon_{ijm} \sim \mathcal{N}(0, \sigma^2)$. The variables x_{ij}, β_{1j}, and β_0 are defined as in the "simple model" description (2.1) in the previous section. However, since each subject had a different number of TCRs measured and varying *TRB* gene usage, we calculated the proportion of TCRs from each candidate *TRB* gene allele group, m, to define a weight, W_{im}, for each observation:

$$W_{im} = \frac{N_{im}}{\sum_{m=1}^{M_t} N_{im}}.$$

With this, we solved the following weighted least squares problem for each SNP, TCR feature, and productivity status combination:

$$(\hat{\beta}_0, \hat{\beta}_{1j}, \hat{\gamma}_j) = \underset{\beta_0, \beta_{1j}, \gamma_{\cdot j}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{m=1}^{M_t} W_{im} \cdot \left(\bar{y}_{im} - (\beta_0 + \gamma_{jm} + \beta_{1j} x_{ij})\right)^2 \tag{2.5}$$

using the `lm` function in R.

With each estimate of the j-th SNP effect on the amount of the TCR feature of interest, $\hat{\beta}_{1j}$, generated using the models described above, we quantified the association strength between each SNP and the amount of the TCR feature by testing whether $\hat{\beta}_{1j} = 0$. To do this, we applied a t-test (described in the previous section) using the test statistic (2.3) to obtain each P-value. However, because our condensed dataset contains a total of M_t observations from each subject i, these P-values may be inflated due to intra-subject observations being potentially correlated. Thus, to increase the accuracy of the P-value calculation, for each association P-value below a certain threshold (we chose $P < 5 \times 10^{-5}$), we recalculated the P-value using a clustered bootstrap (with subjects as the sampling unit). To do so, for each bootstrap iterate, we resampled subjects from the condensed dataset with replacement. Using this re-sampled data, we fit the model in (2.5) to estimate each coefficient. We repeated this bootstrap process 100 times and used the resulting 100 coefficient estimates to estimate a standard error for each model coefficient. With this re-calculated standard error of the estimate of the j-th SNP effect on the amount of the TCR feature of interest, $\text{se}(\hat{\beta}_{1j})$, we wanted to test whether $\hat{\beta}_{1j} = 0$ by recalculating the test-statistic, (2.3), and applying a t-test to obtain each "corrected" P-value. As noted in the multiple testing correction methods section, when accounting for multiple testing via Bonferroni correction, we used the entire number of TCR features and SNPs considered (not just those that were sufficiently promising to warrant use of the bootstrap to get a more accurate P-value): This ensures that our correction will not be anti-conservative.

2.3.6 Correcting for population-substructure-related effects

Structure within our SNP genotype data (such as population-substructure-related biases due to ancestry), if present, may produce false positive associations when quantifying the association strength between each SNP and our phenotype of interest. To account for this, we implemented principal component analysis as commonly applied to genome-wide genotype data for population substructure inference. Specifically, we used the PC-AiR algorithm [58] which identifies principal components that capture ancestry while accounting for relatedness in the samples. As such, the top principal components calculated from the genotype data reflect population substructure among the samples. When plotting the proportion of variance explained by each PC, we find that the majority of variability appears to be explained by the top eight PCs (Figure 2.9). This conclusion is supported when plotting each PC score by ancestral group (Figure 2.9). With this, we incorporated the top eight principal components as covariates into our GWAS models described above.

As such, to quantify the association strength between each SNP and TCR feature without conditioning on gene usage as in (2.1), while incorporating principal component terms to correct for population-substructure-related bias due to ancestry, we fit the model:

$$\bar{y}_i = x_{ij} \cdot \beta_{1j} + \beta_0 + \sum_{p=1}^{8} \beta_{2jp} \cdot P_{ip} + \epsilon_{ij} \tag{2.6}$$

where $\bar{y}_i$, x_{ij}, β_{1j}, β_0, and ϵ_{ij} are defined as in (2.1), β_{2jp} is the population-substructure-related bias correction term for SNP j and the p-th principal component, and P_{ip} is the p-th principal component for subject i as calculated above. To estimate each regression coefficient, we solved the following least squares problem for each SNP, TCR feature, and productivity status combination:

$$(\hat{\beta}_0, \hat{\beta}_{1j}, \vec{\hat{\beta}}_{2j}) = \underset{\beta_0, \beta_{1j}, \vec{\beta}_{2j}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\bar{y}_i - (x_{ij} \cdot \beta_{1j} + \beta_0 + \sum_{p=1}^{8} \beta_{2jp} \cdot P_{ip})\right)^2.$$

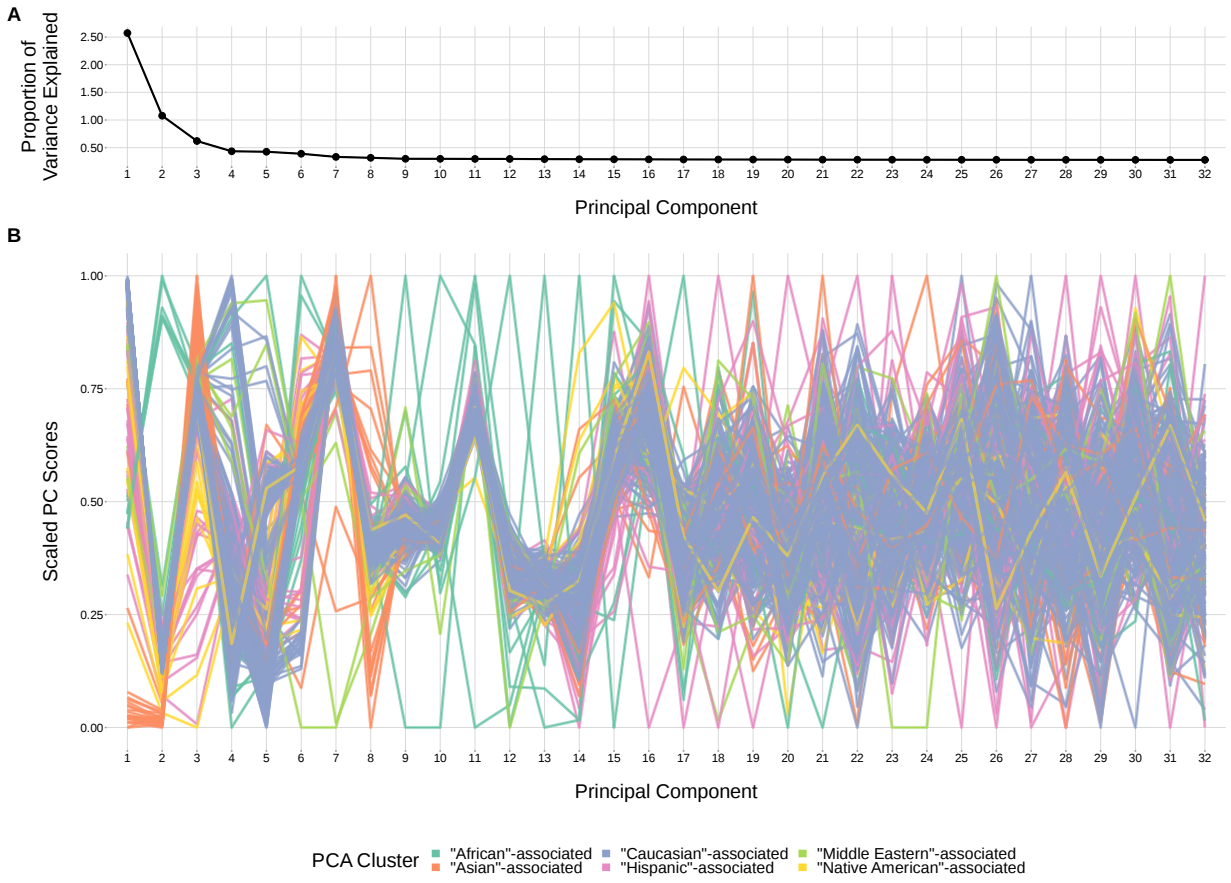

Figure 2.9: The top principal components calculated from genotype data reflect ancestry structure among samples. (**A**) The majority of the ancestry-informative principal component analysis variance is explained by the first 8 principal components. (**B**) The first 8 principal components show distinct separation by PCA cluster. Each colored line represents one of the 398 samples. The first 32 principal components are shown on the X-axis and their scaled component values for each subject on the Y-axis.

Furthermore, to quantify the association strength between each SNP and TCR feature, conditional on gene usage as in (2.4), while incorporating principal component terms to correct for population-substructure-related bias due to ancestry, we fit the model:

$$\bar{y}_{im} = x_{ij} \cdot \beta_{1j} + \beta_0 + \gamma_{jm} + \sum_{p=1}^{8} \beta_{2jp} \cdot P_{ip} + \epsilon_{ijm} \tag{2.7}$$

where $\bar{y}_{im}$, x_{ij}, β_{1j}, β_0, γ_{jm}, and ϵ_{ij} are defined as in (2.4) and β_{2jp} and P_{ip} are defined as in (2.6). Again, to estimate each regression coefficient, we solved the following weighted least squares problem for each SNP, TCR feature, and productivity status combination:

$$(\hat{\beta}_0, \hat{\beta}_{1j}, \hat{\gamma}_j, \vec{\hat{\beta}}_{2j}) = \underset{\beta_0, \beta_{1j}, \gamma_j, \vec{\beta}_{2j}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{m=1}^{M_t} W_{im} \cdot \left(\bar{y}_{im} - (\beta_0 + \gamma_{jm} + \beta_{1j} x_{ij} + \sum_{p=1}^{8} \beta_{2jp} \cdot P_{ip})\right)^2.$$

With these estimates for the population-substructure-corrected j-th SNP effect on the amount of the TCR feature of interest, $\hat{\beta}_{1j}$, we calculated a P-value using the methods described in the methods section for each model type.

2.3.7 Multiple testing correction for associations

For each TCR feature (i.e. extent of trimming, number of N-insertions, etc.), we considered the significance of associations using a Bonferroni-corrected threshold. To establish each threshold, we corrected for each TCR feature subtype (i.e. V-gene trimming, J-gene trimming, etc. for the TCR trimming feature), the two TCR productivity types (productive and non-productive), and the total number of SNPs tested. When considering associations in the whole-genome context, we corrected for the approximately 6.5 million SNPs (remaining after filtering). When considering associations in a gene-level context, we corrected for the number of SNPs within 200 kb of the gene of interest. For the validation analysis, we con-

sidered associations in a SNP-level context and did not correct for multiple SNPs. However, for the validation analysis, we considered the significance of associations within both TCRα and TCRβ chains and, thus, corrected the significance threshold accordingly.

2.3.8 Ancestry-informative PCA cluster classification

In order to correct for population-substructure-related biases due to ancestry in our GWAS analyses, we used ancestry-informative principal component analysis. The original genotyping dataset [57] contained self-reported ancestry. However, a number of subjects did not self-report ancestry in the original data collection. Further, for some subjects, their self-reported ancestry was discordant with clusters observed in a principal component analysis. Therefore, for analysis purposes, we used the minimum covariance determinant method [71, 72] with the original self-identified labels to group the subjects into six ancestry-informative PCA clusters: "African"-associated (8), "Asian"-associated (23), "Caucasian"-associated (322), "Hispanic"-associated (30), "Middle Eastern"-associated (5), and "Native American"-associated (10).

2.3.9 Implementation and code

R code implementing the genome-wide association inferences described here is available at `https://github.com/phbradley/tcr-gwas`. The following tools were especially helpful: `data.table` [73], `tidyverse` [74], `doParallel` [75], `SNPRelate` [76], `GWASTools` [77], `GENESIS` [78], and `cowplot` [79]

Chapter 3

STATISTICAL INFERENCE REVEALS THE ROLE OF LENGTH, GC CONTENT, AND LOCAL SEQUENCE IN V(D)J NUCLEOTIDE TRIMMING

Cells throughout the body regularly present protein fragments, known as antigens, on cell-surface molecules called major histocompatibility complex (MHC). Receptors on the surface of T cells can bind to these MHC-bound antigens, recognize them, and, if necessary, initiate an immune response. For an individual to be capable of defending against a wide array of potential foreign pathogens, they somatically generate a massive diversity of T cell receptors (TCRs) through a random process called V(D)J recombination. After generation, TCRs undergo a selection process to ensure proper expression, MHC recognition, and limited autoreactivity. The collection of TCRs in an individual comprises their TCR repertoire.

The majority of human T cells express α-β receptors that consist of an α and a β protein chain. During the V(D)J recombination process of the β chain, a single V-, D-, and J-gene are randomly chosen from a pool of V-gene, D-gene, and J-gene segments within the germline T cell receptor beta locus over a series of steps. To begin this process, the recombination activating gene (RAG) protein complex aligns a randomly chosen D- and J-gene, removes the intervening chromosomal DNA between the two genes, and forms a hairpin loop at the end of each gene [21–23]. Each hairpin loop is then nicked open, typically in an asymmetrical fashion, by the Artemis:DNA-PKcs protein complex [29, 31]. This asymmetrical hairpin

opening creates a single-stranded DNA overhang at the end of both genes that, depending on the location of the hairpin nick, may contain P-nucleotides (short palindromes of gene terminal sequence) [26, 27, 29–31]. The most dominant hairpin opening position leads to a single-stranded 3' overhang that is 4 nucleotides in length (2 nucleotides of which are P-nucleotides) [31]. From here, nucleotides may be deleted from each gene end through an incompletely-understood mechanism suggested to involve Artemis [2, 25, 27, 30, 32–34, 80, 81]. This nucleotide trimming can remove traces of P-nucleotides [26, 47]. Non-template-encoded nucleotides, known as N-insertions, can also be added to each gene end by the enzyme terminal deoxynucleotidyl transferase (TdT) [35–37]. Once the nucleotide addition and deletion steps are completed, the gene segments are paired and ligated together [32]. From here, the process is repeated between a random V-gene and this combined D-J junction to complete the TCRβ chain. A similar TCRα chain recombination then proceeds, though without a D-gene, to complete the α-β TCR. Other adaptive immune receptor loci, such as *TRG*, *TRD*, and all *IG* loci, also undergo V(D)J recombination during the development of γ-δ T cells and B cells, respectively.

Junctional diversity created by the deletion and non-templated insertion of nucleotides during V(D)J recombination contributes substantially to the resulting diversity of the TCR repertoire. Small variations in gene sequence have been shown to lead to large changes in the extent of nucleotide deletion [25–27, 30]. For example, sequences with high AT content suffer greater nucleotide loss than sequences with high GC content [26]. These findings are suggestive of a nuclease that either binds an AT-rich sequence motif or requires an AT-specific structure (e.g. a sequence that breathes [82]), however, further work is required to quantify this mechanistic preference.

The Artemis protein is often regarded as the main nuclease involved in V(D)J recombination [32, 80, 81]. Artemis is a member of the metallo-β-lactamase family of nucleases [28] and is widely regarded as a structure-specific nuclease as opposed to a nuclease that binds specific

DNA sequences [80, 81, 83, 84]. Members of this family are characterized by their conserved metallo-β-lactamase and β-CASP domains and their ability to nick DNA or RNA in various configurations [85, 86]. Alone, Artemis possesses intrinsic 5'-to-3' exonuclease activity on single-stranded DNA [63]. On double-stranded DNA, Artemis, in complex with DNA-PKcs, has endonuclease activity on 5' and 3' DNA overhangs and on DNA hairpins [29, 31, 87]. It has been proposed that the Artemis:DNA-PKcs complex binds single-stranded-to-double-stranded DNA boundaries prior to nicking [29, 31, 81, 83]; for blunt DNA ends, previous work has concluded that sequence-breathing is required to achieve this structural configuration prior to Artemis action [80]. Further, Artemis, in complex with XRCC4-DNA ligase IV, has additional endonuclease activity on 3' DNA overhangs and preferentially nicks one nucleotide at a time from the single-stranded 3' end [46, 88]. Despite these diverse nucleolytic functions, the extent of involvement and exact mechanism of action for the Artemis protein during the nucleotide trimming step of V(D)J recombination, and how it relates to observed sequence-dependent changes in trimming [25–27, 30], has yet to be fully understood.

While molecular experiments using model organisms have been essential for establishing the current mechanistic understanding of the nucleotide trimming process, studies in humans have been limited. Statistical inference on high-throughput repertoire sequencing data sets allows for exploration of the in-vivo V(D)J recombination mechanism outside of model organisms. In particular, analysis of trimming in high-throughput data sets should lead to insights about the natural underlying process, in the same way that analysis of large data sets has led to insight into the process of somatic hypermutation. There, researchers have found quite significant connections between local sequence identity and mutation patterns, leading to a rich literature [89–97].

In contrast, we are only aware of one statistical analysis connecting sequence identity to trimming lengths [15]. This one existing analysis [15] has shown that a simple position-weight-matrix style model does a surprisingly good job of predicting the distribution of

trimming lengths for a variety of V-genes. However, while this trimming model has good model fit and predictive accuracy, it is limited by the assumption that the trimming mechanism relies solely on a sequence motif and, as such, is not designed in a way that allows us to explore alternative hypotheses.

In this chapter, I explore the sequence-level determinants of nucleotide trimming during V(D)J recombination using statistical inference on high-throughput TCRβ repertoire sequencing data [50]. With the goal of informing our mechanistic understanding in a quantitative way, we have designed a flexible probabilistic model of nucleotide trimming that allows us to explore various sequence-level features. Our results show that trimming probabilities are highest for DNA positions near the end of the sequence that contain high GC content upstream, quantifying the role of sequence-breathing dynamics in the trimming process. We also see evidence of a sequence motif that appears to get preferentially trimmed, independent of possible sequence-breathing effects. As such, we can predict trimming probabilities most accurately using a model that includes features for local sequence context, length, and GC nucleotide content in both directions of the wider sequence. We show that this model has high predictive accuracy for V- and J-gene sequences from an independent TCRβ-sequencing data set, and also extends well to TCRα, TCRγ, and IgH sequences. Further, we demonstrate that genetic variations within the gene encoding the Artemis protein that were previously-identified as being associated with increasing the extent of trimming [2] are also associated with changes in several model coefficients.

3.1 Results

3.1.1 Training data description

We worked with TCRβ-immunosequencing data representing 666 individuals [50]. V(D)J recombination scenarios were assigned to each sequence from each individual using the IGoR

software which is designed to learn unbiased V(D)J recombination statistics from immune sequence reads [11]. Using these V(D)J recombination statistics, IGoR output a list of potential recombination scenarios with their corresponding likelihoods for each TCRβ-chain sequence in the training data set. We annotated each sequence with a single V(D)J recombination scenario by sampling from these potential scenarios according to the posterior probability of each scenario (see Methods for further details).

Annotated TCR sequences can be separated into two categories: "productive" rearrangements which code for a complete, full-length protein and "non-productive" rearrangements which do not. Non-productive sequences are generated when the V(D)J recombination process produces a sequence that is either out-of-frame or contains a stop codon. Each T cell contains two loci which can undergo the V(D)J recombination process. When the first recombination fails to generate a functional receptor (creating a non-productive sequence), followed by a successful rearrangement on the T cell's second chromosome (a productive sequence), the non-productive rearrangement can be sequenced as part of the repertoire. Non-productive sequences do not generate proteins that undergo functional selection in the thymus, and their recombination statistics should reflect only the V(D)J recombination generation process [15, 19, 20]. In contrast, the recombination statistics of productive sequences should reflect both V(D)J recombination generation and functional selection. Because we are interested in nucleotide trimming during the V(D)J recombination generation process, prior to selection, we only include non-productive sequences in our training data set. Further, because V-gene sequences within the *TRB* locus contain more sequence variation than D- and/or J-genes, we only include V-gene sequences in our training data set.

3.1.2 Replicating a previous model of nucleotide trimming

The extent of nucleotide trimming varies substantially from gene to gene [15, 25, 27, 30]. Previous work has identified an interesting impact of sequence features, such as sequence

nucleotide context, on trimming probabilities using a position-weight-matrix-style (PWM) model [15]. To our knowledge, this is the only model that takes nucleotide sequence identity into account when predicting trimming probabilities. Specifically, this model leverages a "trimming motif" containing two nucleotides 5' of the trimming site and four nucleotides 3' of the trimming site to predict the probability of trimming at a given site. It was designed and trained using sequencing data from just nine individuals [15], and has surprisingly good model fit and predictive accuracy across many V-genes despite its simplicity. Using a different, and much larger, repertoire sequencing data set, we have trained this PWM model and replicated previous work (Figure C.1). We will refer to this model as the *2x4 motif* model. It is important to note that this PWM model is not the primary model described in Murugan et. al 2012 [15], but again is the only one that relates nucleotide identity to trimming.

3.1.3 Model set-up overview

While the *2x4 motif* model has good predictive accuracy and model fit [15], it is limited by its assumption that the trimming mechanism relies solely on a sequence motif. Here, we have generalized this PWM model to a model that allows for arbitrary sequence features, and trained each new model using conditional logistic regression (see Materials and Methods). With this set up, we were able to evaluate the relative importance of new mechanistically-interpretable features for predicting trimming probabilities. Specifically, we designed features to measure the effects of DNA-shape, length, and GC nucleotide content in both directions of the wider sequence on the probability of trimming at a given position in a gene sequence.

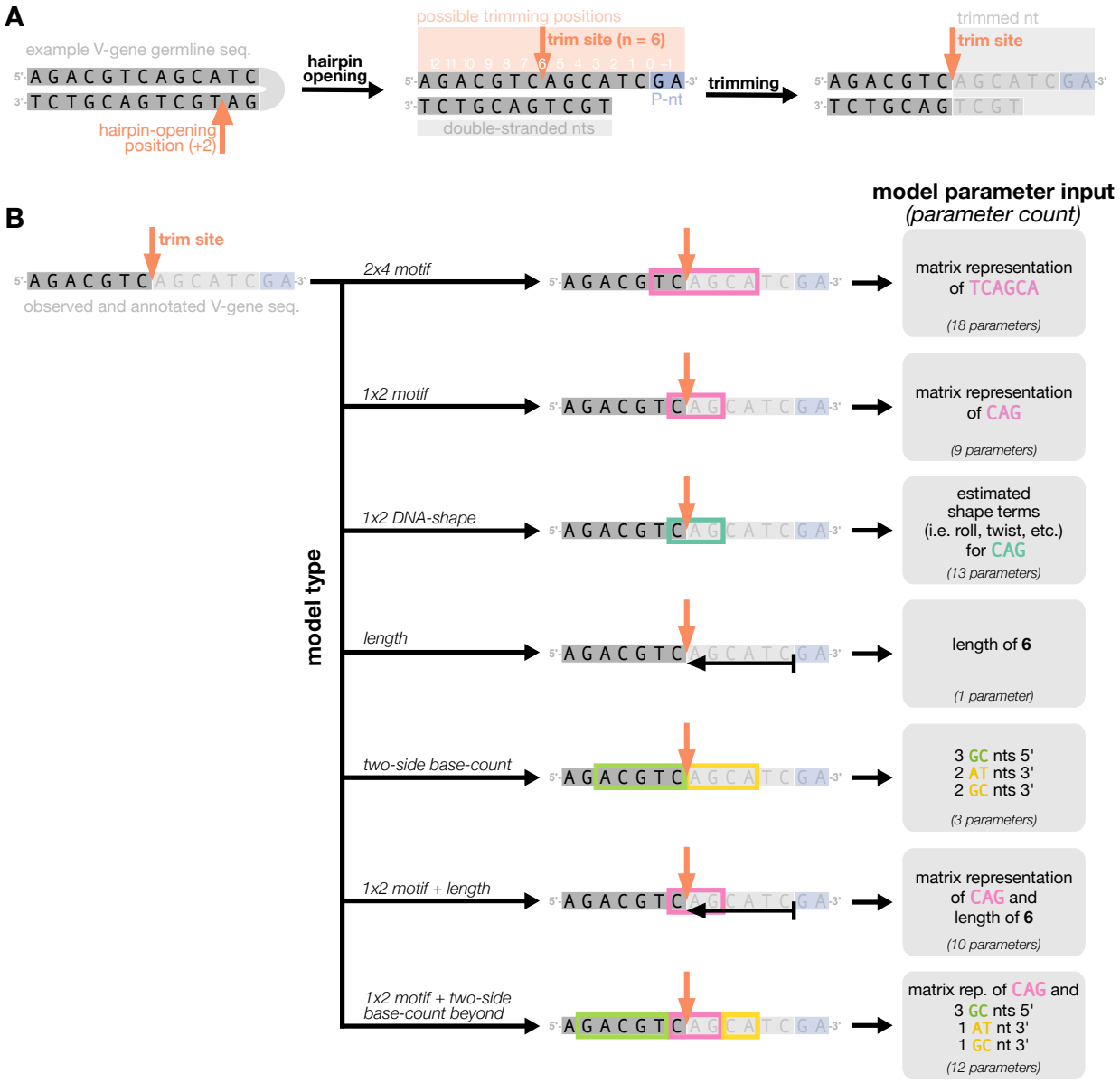
A
example V-gene germline seq.
hairpin opening
hairpin-opening position (+2)
possible trimming positions
trim site (n = 6)
P-nt
double-stranded nts
trimming
trimmed nt
trim site
B
trim site
observed and annotated V-gene seq.
model type
2x4 motif
1x2 motif
1x2 DNA-shape
length
two-side base-count
1x2 motif + length
1x2 motif + two-side base-count beyond
model parameter input
(parameter count)
matrix representation of TCAGCA
(18 parameters)
matrix representation of CAG
(9 parameters)
estimated shape terms (i.e. roll, twist, etc.) for CAG
(13 parameters)
length of 6
(1 parameter)
3 GC nts 5'
2 AT nts 3'
2 GC nts 3'
(3 parameters)
matrix representation of CAG and length of 6
(10 parameters)
matrix rep. of CAG and
3 GC nts 5'
1 AT nt 3'
1 GC nt 3'
(12 parameters)

Figure 3.1: Overview of how a sequence is transformed into features for regression. (**A**) As described, during the early stages of V(D)J recombination between two genes, the hairpin of each gene is opened; here, we are showing this hairpin-opening step for a single arbitrary V-gene. The most common hairpin-opening position leads to a 4-nucleotide-long single stranded overhang (two nucleotides of which are considered P-nucleotides, as shown in purple). From here, each gene can undergo nucleotide trimming. In this example, the V-gene is trimmed back 6 nucleotides. (**B**) All models were trained with non-productive V-gene sequences whose trimming positions were inferred during a sequence annotation step. For our model parameterization, we only consider the top strand (5'-to-3') of the observed sequence. Here, the sequence features parameterized for each model type are shown for the example sequence from (A). The pink boxes surround nucleotides included in the matrix representation of *motif* features. The turquoise boxes surround nucleotides used to estimate and parameterize *DNA-shape* features (see Appendix D for further details). The green boxes surround nucleotides included in the counts of GC nucleotides 5' of the trimming site; in our actual models, we count nucleotides within a 10 nucleotide window (a 5 nucleotide window is shown in the figure). Because this window size is fixed, we do not need to include an additional parameter for AT nucleotide count 5' of the trimming site (since it is already indirectly modeled). The yellow boxes surround double-stranded nucleotides included in the counts of AT and GC nucleotides 3' of the trimming site. These raw 3' nucleotide counts also indirectly parameterize length; as such, we never include both *length* and *two-side base-count* parameters in the same model. In addition to the models shown in the figure, we also evaluated a *null* model which does not contain any parameters.

We parameterize each of these features as follows. An example of how an arbitrary V-gene sequence is transformed into features for modeling is shown in Figure 3.1. To parameterize DNA-shape, we used previously developed methods [98, 99] to estimate various DNA-shape values (i.e. roll, twist, electrostatic potential, minor groove width, etc.) for each single-nucleotide position within a sequence window surrounding the trimming site. To parameterize length, we measure the sequence-independent distance from the end of the gene (i.e. the number of nucleotides from the 3'-end of the sequence) as an integer-valued variable. We parameterize GC nucleotide content using the raw counts of AT and GC nucleotides on both sides of the trimming site (the *two-side base-count*). By using raw nucleotide counts, this measure also serves to parameterize length. Because AT nucleotides have a greater

potential for sequence-breathing compared to GC nucleotides within a sequence [100], these *two-side base-count* terms may be serving as a proxy for the capacity of a sequence to breathe. As such, because sequence-breathing potential is only relevant for nucleotides that are paired, we do not include the nucleotides within the 3' single-stranded-overhang when counting 3' AT and GC nucleotides (see Appendix D).

With these features, we designed models containing various feature combinations (Figure 3.1B). Collectively, these models allow us to explore other possible sequence-level determinants of nucleotide trimming, in addition to the previously proposed [15] "trimming motif" hypothesis. We trained each model using the V-gene training data set described above (see Materials and Methods for further model training details), and evaluated performance using a suite of different held-out data groups (Figure 3.2). Specifically, to evaluate model fit, we computed the expected per-sequence conditional log loss of each model using the full V-gene training data set.

To evaluate model generalizability, we computed the expected per-sequence conditional log loss using the following held-out groups:

- many random, held-out subsets of the V-gene training data set
- held-out subsets of the V-gene training data set containing groups of V-genes defined to be the "most-different" from all other genes using either the terminal sequences (last 25 nucleotides of each sequence) or the full gene sequences
- the full J-gene data set

For each of these held-out group analyses, each model was re-trained using the full V-gene training data set with the held-out group-of-interest removed (see Materials and Methods and Appendix D for further details) prior to computing the loss. A lower expected per-sequence conditional log loss indicated better model fit and/or model generalizability. Following this model evaluation, we validated a subset of the models by using the model coefficients from the

previous TCRβ V-gene training run and computing the expected per-sequence conditional log loss of the model using several independent testing data sets (Figure 3.2).

3.1.4 Local sequence context, length, and GC nucleotide content in both directions of the wider sequence, together, accurately predict the trimming probabilities of a given V-gene sequence

In an effort to capture the complex underlying biochemistry of the deletion process, we trained models containing various combinations of sequence-level feature types (Figure 3.1B) and evaluated their ability to accurately predict V-gene trimming probabilities. With this approach, we found that a model containing parameterizations of local sequence context, length, and GC nucleotide content in both directions of the wider sequence (the *1x2 motif + two-side base-count beyond* model) had the best model fit and generalizability across different data sets (Figure 3.3). This model contains a *1x2 motif*, including one nucleotide position 5' of the trimming site and two nucleotide positions 3' of the trimming site within the trimming window, and includes only bases beyond this trimming window in the AT and GC *two-side base-count* terms (Figure 3.1). Despite containing fewer total parameters than the original *2x4 motif* model [15] (12 parameters compared to 18 parameters), the *1x2 motif + two-side base-count beyond* model had better predictive accuracy (Figure 3.4A and Figure C.3).

We considered the significance of the inferred model coefficients using a Bonferroni-corrected significance threshold of 0.0033 (corrected for the total number of model coefficients). With this threshold, we found that many of the inferred model coefficients were significant and quantified mechanistic patterns. Each coefficient represents the change in log10 odds of trimming at a given site resulting from an increase in the feature value, given that all other features are held constant. Within the nucleotides immediately surrounding the trimming site, bases 5' of the trimming site have a slightly stronger effect on the trimming probability than bases 3' of the trimming site (Figure 3.4B). Specifi-

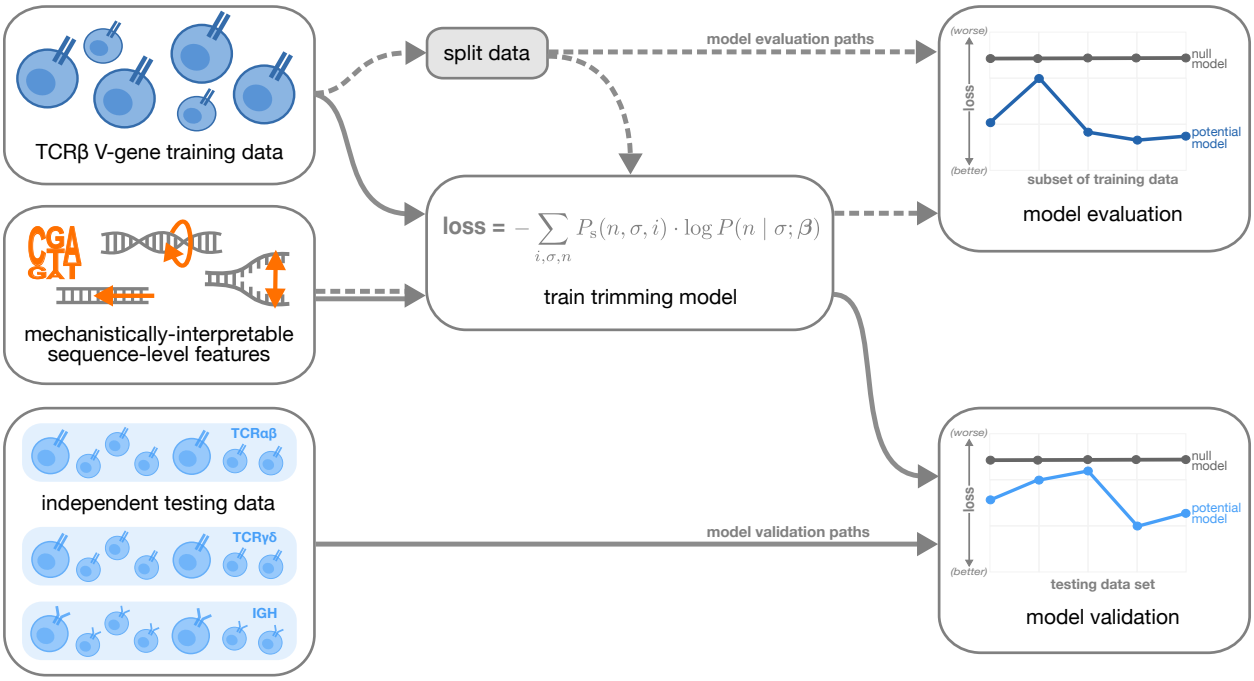

Figure 3.2: Overview of analysis strategy. The TCRβ V-gene training data was used to train each trimming model containing various combinations of sequence-level features (Figure 3.1) by minimizing the associated loss function. The loss function is given by a sum across individuals i, genes σ, and trimming lengths n of the sampling probability of each observation P_s multiplied by the gene-specific trimming probability predicted by a model with β parameters (see Methods for further details). Each potential model first underwent a "model evaluation" stage (shown by the dashed lines) during which the model performance was evaluated using various subsets of the training TCRβ V-gene data set. Once all models were evaluated, a subset of the potential models continued on to the "model validation" stage (shown by the solid lines) during which the performance of the model coefficients from the previous TCRβ V-gene training run were validated using several independent testing data sets including TCRβ, TCRα, TCRγ, and IgH sequences. At each stage, the performance of each model was compared to a null model (containing zero parameters, see Methods).

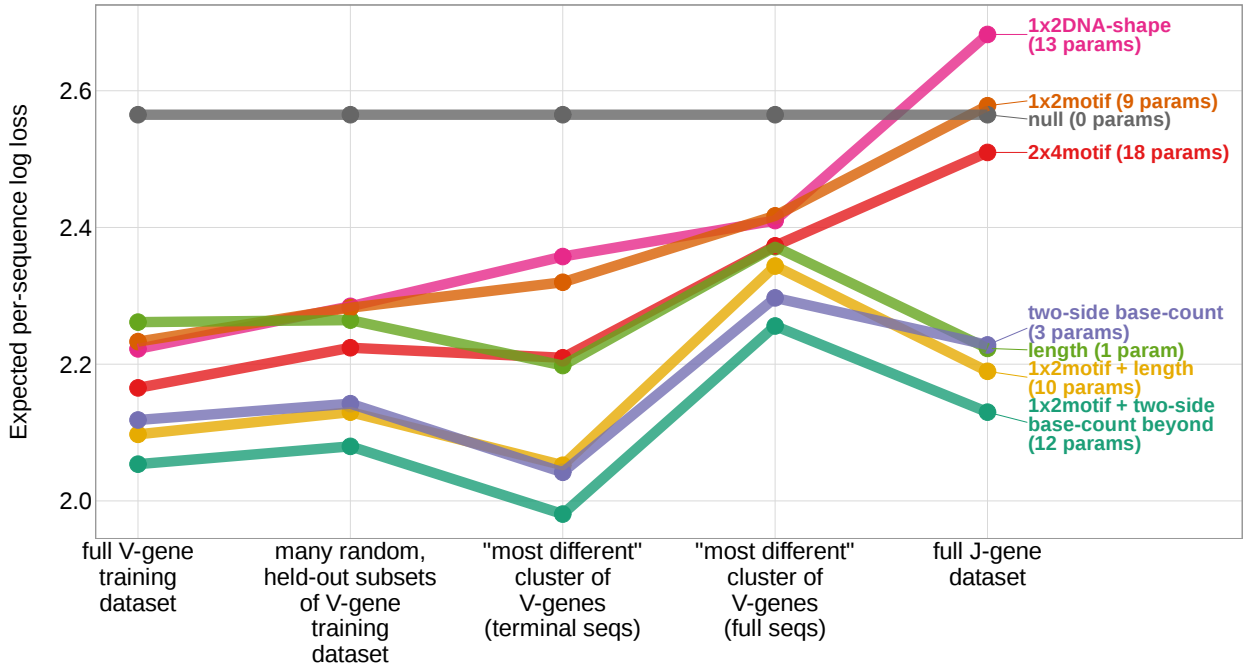

Figure 3.3: Expected per-sequence conditional log loss computed for various models using the full V-gene training data set, many random, held-out subsets of the V-gene training data set, a held-out subset of the V-gene training data set containing a group of V-genes defined to be the "most-different" using the terminal sequences (last 25 nucleotides of each sequence), a held-out subset of the V-gene training data set containing a group of V-genes defined to be the "most-different" using the full gene sequences, and the full J-gene data set. Each model was trained using the full V-gene training data set with the held-out group or "most-different" group (if applicable) removed (see Materials and Methods and Appendix D). Lower expected per-sequence log loss corresponds to better a model fit. The *1x2 motif + two-side base-count beyond* model has the best model fit and generalizability across all data sets.

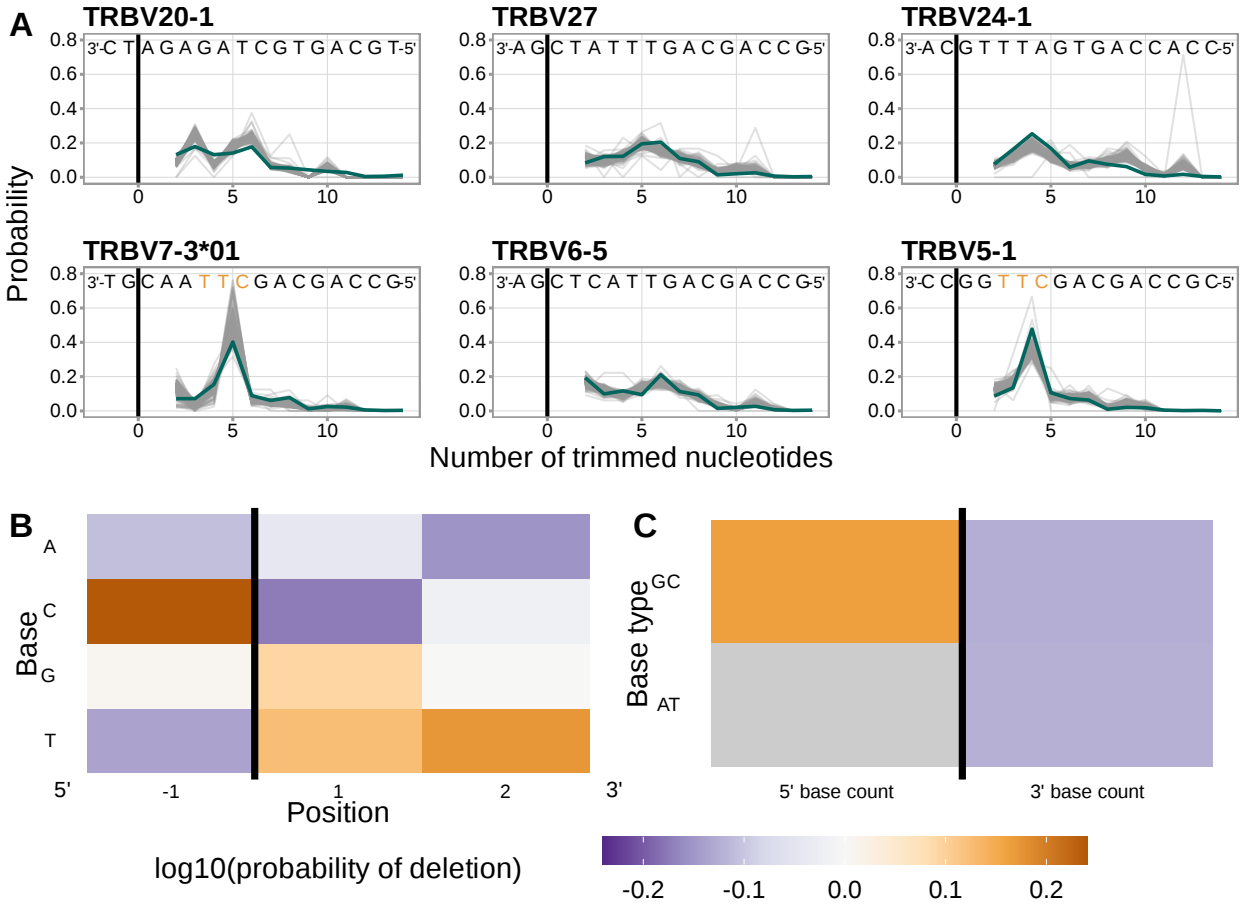

Figure 3.4: Inferred coefficients and performance of the *1x2 motif + two-side base-count beyond* model. (**A**) Inferred trimming profiles using the *1x2 motif + two-side base-count beyond* model have good predictive accuracy overall; here we show the inferred trimming profiles (in blue) for the most frequently used V-genes. Gene-specific trimming profiles for each individual in the training data set are shown in gray. The sequence context with the highest probability of trimming (3'-TTC-5' or 3'-TGC-5') from (B and C) is highlighted in orange. (**B**) Position-weight-matrix of the local sequence context dependence of V-gene trimming probabilities consisting of 1 nucleotides 5' of the trimming site and 2 nucleotides 3' of the trimming site from fitting the *1x2 motif + two-side base-count beyond* model. Positions 5' and 3' of the trimming site have a strong effect on the probability of trimming. (**C**) Inferred *two-side base-count beyond* model coefficients from fitting the *1x2 motif + two-side base-count beyond* model suggest that the count of GC bases 5' of the motif has a strong positive effect on the trimming probability whereas the count of GC and/or AT bases 3' of the motif has a negative effect. The count of AT nucleotides 5' of the motif (shown in gray) was not included in this model. The black vertical line corresponds to the trimming site. Each inferred coefficient is given as the change in log10 odds of trimming at a given site resulting from an increase in the feature value, given that all other features are held constant.

cally, 5' of the trimming site, C nucleotides have a strong positive effect on the trimming probability ($\log_{10}$ coefficient $= 0.2388$) whereas A and T nucleotides have a negative effect ($\log_{10}$ coefficient$_A = -0.108$ and $\log_{10}$ coefficient$_T = -0.137$). In contrast, immediately 3' of the trimming site, G and T nucleotides have a positive effect on the trimming probability ($\log_{10}$ coefficient$_G = 0.093$ and $\log_{10}$ coefficient$_T = 0.125$) whereas C nucleotides have a negative effect ($\log_{10}$ coefficient $= -0.174$). These results suggest a different possible mechanistic pattern than previous *motif*-only models [15] (Figure C.1B). Further, beyond the *1x2 motif* sequence-window, the count of GC nucleotides 5' of the motif (within a 10 nucleotide window) has a strong positive effect on the trimming probability ($\log_{10}$ coefficient $= 0.164$) (Figure 3.4C). The counts of both AT and GC nucleotides 3' of the motif have a strong negative effect on the trimming probability ($\log_{10}$ coefficient$_{AT} = -0.123$ and $\log_{10}$ coefficient$_{GC} = -0.126$). Interestingly, the magnitude of these negative effects are very similar between AT and GC counts. This suggests that the raw number of nucleotides 3' of the motif (e.g. the length) is more important for predicting the trimming probability at a given site compared to the identity of the nucleotides. P-values for each of these coefficients were reported to be smaller than machine tolerance (2.23×10^{-308}). We noted minimal variation in the magnitude of each inferred coefficient even when changing the number of sequences included in the training data set (Figure C.9).

Because we were interested in parameterizing sequence-breathing effects using the *two-side base-count* terms, we only included nucleotides that are considered to be double-stranded after hairpin-opening within each count. In our modeling, we assume that the DNA-hairpin is opened at the $+2$ position, leading to a 4 nucleotide long 3'-single-stranded-overhang (the 2 nucleotides furthest 3' are considered P-nucleotides) [29, 31]. As such, the first 2 nucleotides of the gene sequence can be considered single-stranded, and we do not include them in the *two-side base-count* terms. When we train a model that ignores this distinction, and include all gene sequence nucleotides in the *two-side base-count* terms, we note very similar inferred

coefficients and model fit (Figure C.4). We acknowledge that other hairpin-opening positions may be possible. To explore whether the +2-hairpin-opening-position assumption could be affecting our inferences, we trained the *1x2 motif + two-side base-count beyond* model with other possible hairpin-opening-position assumptions and noted minimal variation in model fit (Figure C.5).

We also evaluated the predictive accuracy of *motif + two-side base-count beyond* models containing different "trimming motif" sizes. We find that models containing a small motif (e.g. a *1x2 motif*) achieve similar predictive accuracy and are more generalizable compared to models containing a larger motif (Figure C.6).

Because the trimming mechanism is thought to be consistent across V, D, and J genes from both productive and non-productive sequences, we were also interested in whether the inferred coefficients for the *1x2 motif + two-side base-count beyond* model would be consistent between the model trained using the non-productive V-gene training data set, a model trained using a non-productive J-gene data set, and a model trained using a productive V-gene data set. As such, we trained a new *1x2 motif + two-side base-count beyond* model using only non-productive J-gene sequences and a separate, new *1x2 motif + two-side base-count beyond* model using only productive V-gene sequences (both sequence sets were from the same cohort of individuals as the V-gene training data set). We found that the inferred coefficients were highly similar between the three models (Figure C.7 and Figure C.8).

When evaluating models containing only a single feature type, we find that the *two-side base-count* model which parameterizes GC nucleotide content on both sides of the trimming site (and, indirectly, length) has the best model fit and generalizability across all held-out groups tested (Figure 3.3). As such, these GC-content features, which are likely parameterizing the capacity for the sequence to breathe, are more predictive of V-gene trimming probabilities than local sequence context or DNA-shape alone. This finding supports previous observations that Artemis may act as a structure-specific nuclease as opposed to a

nuclease that binds specific DNA sequences [80, 81, 83, 84].

3.1.5 Inferred local sequence context coefficients suggest a biological trimming motif

A persistent concern with the *1x2 motif + two-side base-count beyond* model was that the *motif* coefficients could be driven by certain genes, instead of representing an actual gene-segment-wide signal. When comparing the inferred trimming profiles from the *two-side base-count* model to those from the *1x2 motif + two-side base-count beyond* model, we identified a group of V-genes which had drastically lower prediction error when the *1x2 motif* terms were included. These V-genes had a difference in per-gene root mean squared error between the two models that was greater than -0.13 (Figure 3.5A). The genes included in this group were *TRBV5-3*, *TRBV7-3*01*, *TRBV7-3*04*, *TRBV7-4*, *TRBV9*, *TRBV11*, and *TRBV13*. To evaluate whether these genes could be driving the observed *motif* signal, we explored whether the prediction error for these genes changed when they were removed from the model training data set.

In fact, we found that the inferred trimming profiles for these genes still had very low prediction error despite the genes not being included in the model training data set (Figure 3.5B, C), showing the generalizability of these features. The inferred model coefficients from this *1x2 motif + two-side base-count beyond* model fit using the subsetted training data set were highly similar to those from the original model fit using the full training data set. Because genes which are highly-similar sequence-wise to the group of held-out genes could still be present in the training data set and be driving these similarities, we defined a new data set that excluded this larger group of genes. When we repeated the same experiment with this new, more-restricted training data set, we observed similar results (Figure 3.5B, C). As such, both of these experiments provided evidence that the *motif* signal may actually represent a gene-segment-wide sequence motif that appears to get preferentially trimmed, independent of GC-content-related effects.

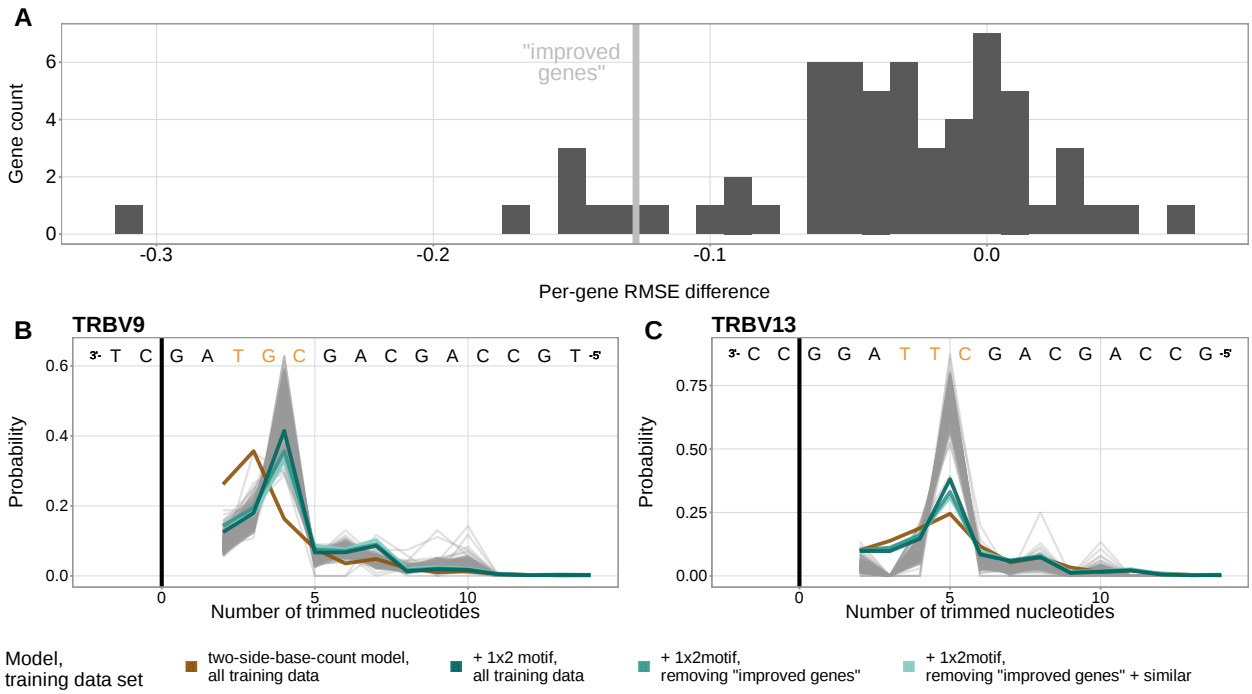

Figure 3.5: The *1x2 motif* coefficients represent a gene-segment-wide trimming motif. (**A**) Distribution of the difference in per-gene root mean squared error (RMSE) between the *1x2 motif + two-side base-count beyond* model and the *two-side base-count* model. V-genes with an RMSE difference less than -0.127 (gray vertical line) were in the lowest 10% of all RMSE differences. These “improved genes” showed a large RMSE improvement when including motif terms in the model. (**B**) Inferred trimming profiles for *TRBV9*, the gene which showed the largest RMSE improvement in (A). *TRBV9* had a RMSE difference of -0.31. (**C**) Inferred trimming profiles for *TRBV13*, the gene which showed the second largest RMSE improvement in (A). *TRBV13* had a RMSE difference of -0.15. The inferred trimming profiles for *TRBV9* and *TRBV13* using models which contain motif terms have very low prediction error even when the genes are not included in the model training data set. Gene-specific trimming profiles for each individual in the training data set are shown in gray. The sequence context with the highest probability of trimming (3’-TTC-5’ or 3’-TGC-5’ from Figure 3.4B) are highlighted in orange.

3.1.6 Trimming-associated variation within the Artemis locus is associated with a change in model coefficients

Using a subset of the V-gene training data set used here, we previously identified a set of single nucleotide polymorphisms (SNPs) within the gene encoding the Artemis protein that are associated with increasing the extent of V- and J-gene trimming [2]. This result suggested that trimming profiles may subtly vary in the context of these SNPs. As such, we were interested in whether these SNPs could be mediating (or serving as a proxy for) a change in the trimming mechanism. To explore this, we worked with paired SNP array and TCRβ-immunosequencing data representing 611 of the original 666 individuals in the V-gene training data set used here. Our previous work [2] used data from only 398 of these individuals, however, the conclusions of that paper held when using this expanded group of 611 individuals in the analysis. With these data, we asked whether the inferred coefficients from the V-gene-specific *1x2 motif + two-side base-count beyond* model varied significantly in the context of the non-coding Artemis-locus SNP (rs41298872) that was found to be most strongly associated with increasing the extent of V-gene trimming in our previous work [2]. As such, we re-defined the model to include an interaction coefficient between the SNP genotype and each model parameter (see Materials and Methods). We then used a Bonferroni-corrected significance threshold of 0.0033 (corrected for the total number of interaction coefficients) to evaluate the significance of each interaction coefficient. For each significant interaction coefficient, we concluded that the corresponding model coefficient varied significantly in the context of the SNP genotype.

Using these methods, we found that several of the *1x2 motif + two-side base-count beyond* model coefficients varied significantly in the context of the Artemis-locus SNP rs41298872 (Figure 3.6). Specifically, we found that 3' of the trimming site, the negative effect of A nucleotides on the trimming odds varied in the context of the SNP for the position immediately 3' of the trimming site ($\log_{10}$ interaction coefficient = 0.006, $P = 0.0006$) and one position

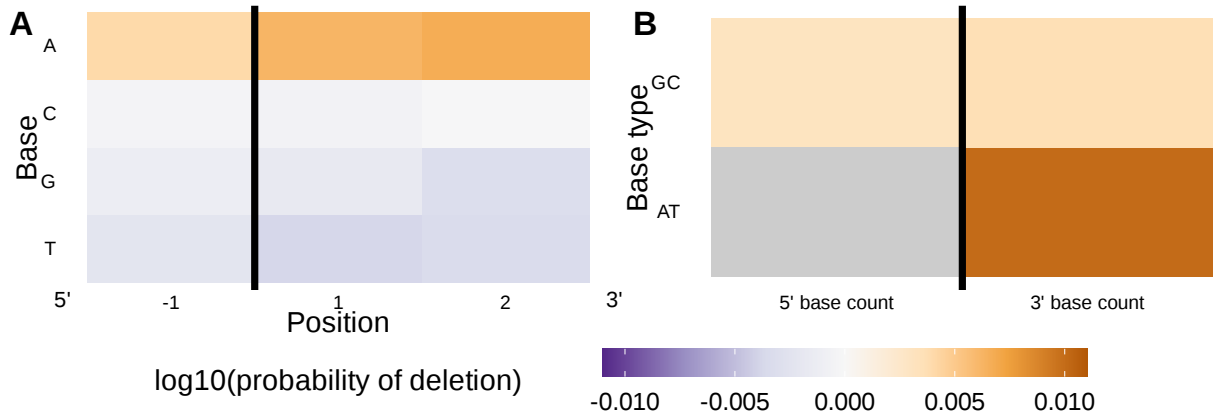

Figure 3.6: Inferred SNP-parameter-interaction coefficients from fitting the *1x2 motif + two-side base-count beyond* SNP interaction model. Note that the inferred coefficients for each main parameter (as shown in Figure 3.4) are not displayed here; only the inferred interaction coefficients between the SNP and each parameter are shown. (**A**) Inferred interaction coefficients between rs41298872 SNP genotype and *motif* parameters for 1 nucleotide position 5' of the trimming site and 2 nucleotide positions 3' of the trimming site. The interaction coefficients between the SNP genotype and the presence of A nucleotides (at all positions 3' of the motif) are significant. This figure is a different representation of the information shown in (A). (**B**) Inferred interaction coefficients between rs41298872 SNP genotype and *two-side base-count beyond* model coefficients. The interaction coefficients between the SNP genotype and the count of AT nucleotides 3' of the motif are significant. The interaction coefficient between the SNP genotype and the count of AT nucleotides 5' of the motif (shown in gray) was not included in this model. The black vertical line corresponds to the trimming site. Each inferred interaction coefficient is given as the change in log10 odds of trimming at a given site resulting from an increase in the feature value and a change in genotype, given that all other features are held constant.

away ($\log_{10}$ interaction coefficient = 0.007, $P = 0.0006$). Further, we found that the negative effect of the count of AT nucleotides 3' of the motif varied strongly in the context of the SNP ($\log_{10}$ interaction coefficient = 0.010, $P = 1.47 \times 10^{-12}$). No other *motif* or *two-side base-count* coefficients were found to significantly vary.

Because the 3'-side *base-count-beyond* terms parameterize both GC nucleotide content and length in their definition, we were interested in whether the significance of the 3'-AT-nucleotide count SNP variation effect was related to GC nucleotide content, length, or both. To do this, we re-defined the 3'-side *base-count-beyond* parameters to be a proportion instead of raw AT/GC nucleotide counts and included an additional *length* term in the model to remove length-related effects from the inferred 3'-side *base-count-beyond* coefficients. Using this new model, we repeated the analysis and found that the *length* coefficient varied significantly in the context of the SNP ($\log_{10}$ interaction coefficient = 0.005, $P = 6.24 \times 10^{-23}$), but the 3'-AT-nucleotide-proportion term did not (Figure C.10). This result is fully consistent with the fact that the Artemis-locus SNP is known to be associated with increasing the extent of trimming (a proxy for length).

3.1.7 Local sequence context, length, and GC nucleotide content in both directions of the wider sequence can also accurately predict the trimming probabilities of a given sequence from other receptor loci

To validate our previously-trained models, we worked with TCRα- and TCRβ-immunosequencing data representing 150 individuals, TCRγ-immunosequencing data representing 23 individuals, and IgH-immunosequencing data representing 9 individuals from three independent validation cohorts. Before analyzing these data, we "froze" our trained model coefficients in git commit 093610a on our repository. In contrast to the training data cohort, these validation cohorts contain different demographics and were each processed using different sequence annotation methods (see Materials and Methods). To explore the potential

effects of using a different sequence annotation method, we re-annotated the TCRβ training data set using the same annotation method as the TCRα-β testing data and found that it had little to no effect on the model fit or performance (Figure C.11).

To evaluate the performance of the *1x2 motif + two-side base-count beyond* model using these testing data, we used the model coefficients from the previous TCRβ V-gene training run and computed the expected per-sequence conditional log loss of the model using each testing data set (TCRβ V-gene sequences, TCRα V-gene sequences, TCRγ V-gene sequences, IgH V-gene sequences, TCRβ J-gene sequences, etc.). We found that the model has high predictive accuracy (i.e. low expected per-sequence conditional log loss) for both non-productive V- and J-gene sequences from the TCRβ testing data set (Figure 3.7). The model also extends well to non-productive V- and J-gene sequences from the TCRα and TCRγ testing data sets and to non-productive V-gene sequences from the IgH testing data set. The model has relatively poor predictive accuracy for non-productive IgH J-gene sequences, however. We noted very similar results when validating model performance using productive V- and J-gene sequences from each testing data set (Figure C.12).

We hypothesized that the weight of the *1x2 motif* and *two-side base-count beyond* model terms may vary across each testing data set. To explore this for each data set, we again used the model coefficients from the previous TCRβ V-gene training run and trained a new two-parameter model containing one coefficient scaling the *1x2 motif* terms and a second coefficient scaling the *two-side base-count beyond* terms (see Materials and Methods). With this approach, we found that the *two-side base-count beyond* terms were dominant compared to the *1x2 motif* terms for every data set (Figure C.13A). The scale coefficient for the *1x2 motif* terms was very small for several of the data sets, especially the IgH data set, indicating only a weak motif-related signal. The sequence motifs that lead to a large increase in trimming probabilities in the model appear at relatively low frequencies within the germline *IGH* genes (Figure C.14), perhaps explaining the weakness of the motif-related signal. When

evaluating the expected per-sequence conditional log loss of these partially re-trained models, we note a small improvement in model fit for each re-trained model compared to the original model (Figure C.13B).

3.2 Discussion

The junctional deletion and insertion steps of the V(D)J recombination process are essential for creating diversity within the TCR repertoire. While the Artemis protein is often regarded as the main nuclease involved in V(D)J recombination, the exact mechanism of nucleotide trimming has yet to be understood in a human system. Using a previously-published high-throughput TCRβ sequencing data set, we designed a flexible probabilistic model of nucleotide trimming that allowed us to explore the relative importance of various sequence-level features. While we recognize that these general model features may not capture the full complexity of the trimming mechanism and establish causation, we were primarily interested in identifying mechanistically-interpretable features which could confirm and extend our current understanding of the nucleotide trimming process. With this framework, we have (1) revealed a set of sequence-level features which can be used to accurately predict trimming probabilities across various adaptive immune receptor loci, (2) shown that length and GC nucleotide content in both directions of the wider sequence are highly predictive of trimming probabilities, quantifying how double-stranded DNA needs to be able to breathe for trimming to occur, (3) identified a sequence motif that appears to get preferentially trimmed, independent of length- and GC-content-related effects, and (4) demonstrated that a genetic variant within the gene encoding the Artemis protein is associated with changes in several model coefficients.

Specifically, we find that a model containing parameterizations of both local sequence context, length, and GC nucleotide content in both directions of the wider sequence can

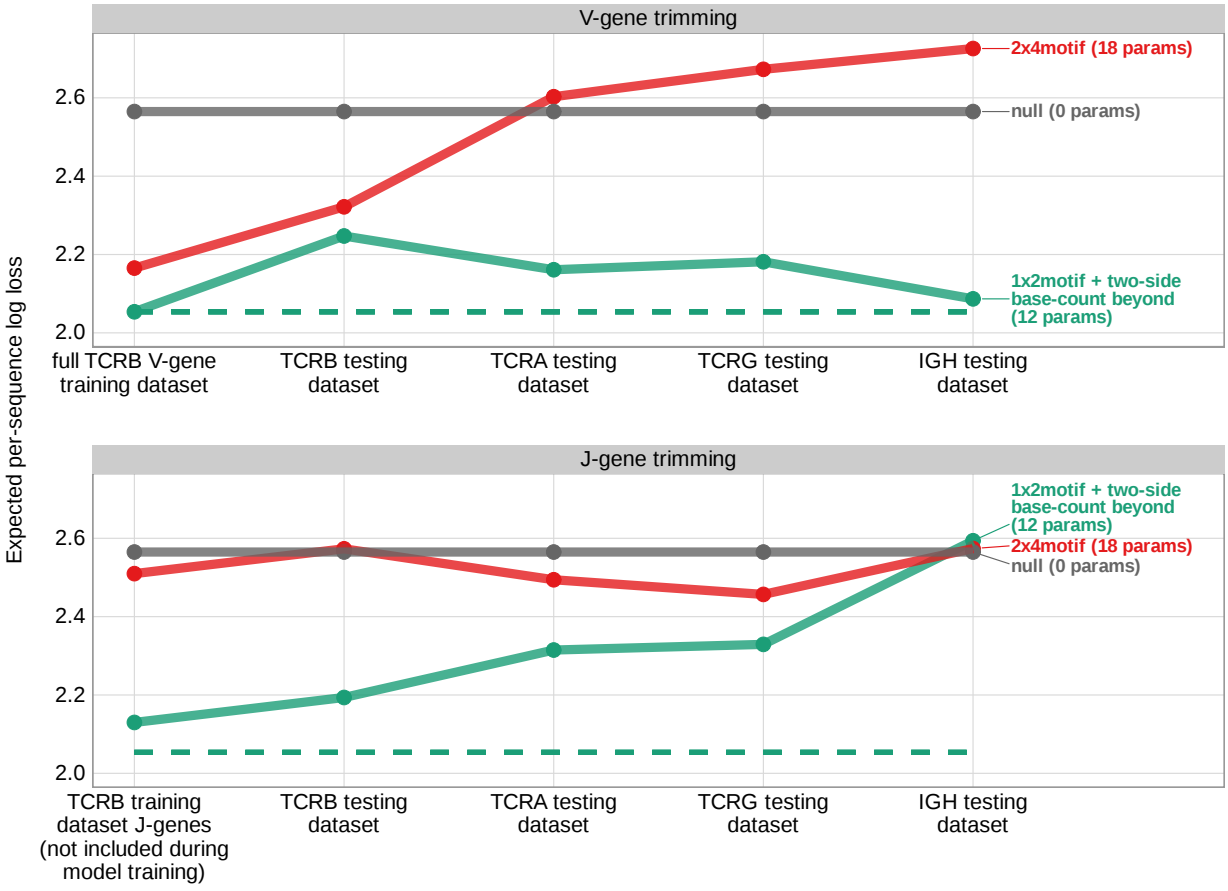

Figure 3.7: Expected per-sequence conditional log loss computed for various models using the TCRβ V-gene training data set and non-productive V- and J-gene sequences from several independent testing data sets. Each model was trained using the full non-productive TCRβ V-gene training data set. Lower expected per-sequence log loss corresponds to a better model fit. The *1x2 motif + two-side base-count beyond* model has the best model fit and generalizability across all testing data sets. The horizontal dashed line corresponds to the expected per-sequence log loss of the *1x2 motif + two-side base-count beyond* model computed for V-gene trimming using the non-productive TCRβ V-gene training data set.

most accurately predict the trimming probabilities of a given TCRβ gene sequence. In addition to having fewer parameters, this model also had better predictive accuracy than a previously proposed sequence context model [15]. Models containing other sequence-level parameters such as DNA-shape and length also had relatively worse predictive accuracy. The *TR* and *IG* V(D)J recombination processes, including trimming profiles, have previously been suggested to vary substantially across individuals [2, 101]. Here, our results support a universal, sequence-based trimming mechanism underlying this variation across *TR* and *IG* loci in humans. Specifically, in addition to TCRβ sequences, we find that local sequence context, length, and GC nucleotide content in both directions of the wider sequence can be used to accurately predict trimming probabilities across TCRα, TCRγ, and IgH sequences. For all of these loci, we find that length and GC nucleotide content are relatively more important than local sequence context terms for making accurate model predictions.

The Artemis protein, in complex with DNA-PKcs, is responsible for opening the DNA hairpin during the early steps of V(D)J recombination to generate a four-nucleotide-long 3'-single-stranded overhang at the end of each gene, and has been suggested to continue on to trim nucleotides from this resulting DNA structure [2, 25, 27, 30, 32–34, 80, 81]. The Artemis protein, with and without DNA-PKcs, has been shown to bind single-stranded-to-double-stranded DNA boundaries prior to nicking DNA [29, 80, 81, 83]. While the single-stranded overhang created during hairpin-opening may create a natural single-stranded-to-double-stranded DNA substrate for Artemis binding near the end of the gene sequence, we find that many trimming events occur further into the double-stranded gene sequence. Indeed, previous in-vitro DNA nuclease assays involving Artemis have shown that sequence-breathing dynamics are often required to generate a transient single-stranded-to-double-stranded DNA substrate prior to Artemis action [80]. Using our model of nucleotide trimming, we have shown that trimming probabilities are highest for DNA positions closer the end of the sequence. Because these DNA positions have fewer double-stranded nucleotides on the 3'-side

of the trimming site, they may have more capacity for sequence-breathing. On the 5'-side of the trimming site, we find that having a larger number of G-C nucleotides, and perhaps less sequence-breathing capacity, increases the trimming probability. Perhaps this breathing transition can create a transient single-stranded-to-double-stranded DNA substrate that is suitable for Artemis to bind and trim. As such, this finding quantifies sequence-breathing effects that were previously identified through in-vitro DNA nuclease assay studies involving Artemis [80].

Independent of GC-content-related effects, we have also identified a gene-segment-wide sequence motif that appears to get preferentially trimmed. This motif is suggestive of sequence-specific nucleolytic activity, however, Artemis is widely regarded as a structure-specific nuclease as opposed to a nuclease that binds specific DNA sequences [80, 81, 83, 84]. This suggests that either (1) Artemis actually does possess some ability to recognize specific nucleotides, (2) the observed sequence motif is serving as a proxy for DNA structure induced by the motif, or (3) another nuclease, in addition to Artemis, is responsible for the sequence-specific trimming we observe. However, because the strength of this sequence motif signal varied across receptor loci, further work will be required to explore its mechanistic basis and presence.

We found that several model coefficients related to local sequence context, length, and GC nucleotide content in both directions of the wider sequence varied significantly in the context of the non-coding Artemis-locus SNP rs41298872. We previously identified this Artemis locus SNP as being associated with increasing the extent of TCRβ V- and J-gene trimming [2]. While many previous studies have reported a high consistency of TCRβ trimming profiles across individuals [11, 15, 17], our results begin to explore how the trimming mechanism may vary across individuals in the context of Artemis genetic variation. We reported that trimming probabilities decrease as the number of double-stranded nucleotides 3' of the trimming site increases. In the context of the SNP rs41298872, we found that as the

number of double-stranded AT nucleotides 3' of the trimming site increases, the trimming probabilities do not decrease as quickly. This suggests that individuals homozygous (or heterozygous) for rs41298872 may be more capable of trimming at positions that have a larger number of double-stranded nucleotides 3' of the trimming site, especially if the additional nucleotides are AT bases. This may be possible if, for example, rs41298872 increases Artemis expression. If there is more Artemis available, then trimming at less optimal positions (i.e. positions further into the sequence which have less breathing) may be possible. Additional work will be required to define the relationship between rs41298872 genotype and Artemis expression.

We also identified several local sequence context coefficients that varied in the context of rs41298872, however, their mechanistic interpretation remains unclear. Earlier, we noted that A nucleotides 3' of the trimming site have a negative effect on the trimming probability while T nucleotides have a strong positive effect. In the context of rs41298872, we found that the magnitude of the negative effect of 3' A nucleotides on the trimming probability was reduced. This may suggest that individuals homozygous (or heterozygous) for rs41298872 may trim in a less motif-dependent fashion, and are instead more reliant on sequence openness 3' of the trimming site. In this way, having A or T nucleotides 3' of the trimming site would create a more open local sequence for trimming.

There are several key limitations of our approach which are intrinsic to the use of adaptive immune receptor repertoire data. First, we have used trimming statistics from non-productive rearrangements as a means of studying the nucleotide trimming process in the absence of selection. Non-productive sequences can be sequenced as part of the repertoire when they are present within a cell expressing a productive rearrangement that survived the selection process. While we are not aware of a mechanism through which non-productive and productive rearrangements within a single cell could be correlated, we also acknowledge that the repertoire of non-productive rearrangements may be an imperfect proxy for a pre-

selection repertoire. However, as is common in the literature [11, 15, 17, 19, 20], we assume that the two recombination events are independent and that the non-productive rearrangements reflect the statistics of the repertoire prior to selection. Next, because many V(D)J rearrangement scenarios can give rise to the same final nucleotide sequence, possible error related to the annotation of each sequence may have restricted our ability to model the actual trimming distributions of each gene. Although we cannot rule out some effect of incorrect sequence annotation on our model inferences, we found that the exact sequence annotation method used, including sampling from the posterior distribution of rearrangement events, had little to no effect on the model fit or performance.

In summary, we have found that local sequence context, length, and the GC nucleotide content in both directions of the wider sequence can accurately predict the trimming probabilities of *TR* and *IG* gene sequences. These results refine our understanding of how nucleotides are trimmed during V(D)J recombination. The sequence-level features identified here lay the groundwork for further exploration into the trimming mechanism and how it may vary across individuals. Such insights will provide another step towards understanding how V(D)J recombination generates diverse receptors and supports a powerful, unique immune response in humans.

3.3 Methods and Materials

3.3.1 Data details

Training data set: TCRβ repertoire sequence data for 666 healthy bone marrow donor subjects was downloaded from the Adaptive Biotechnologies immuneACCESS database using the link provided in the original publication [50]. V(D)J recombination scenarios were assigned to each sequence for each individual using the IGoR software (version 1.4.0) [11] as follows. The IGoR software can learn unbiased V(D)J recombination statistics from immune

sequence reads. Using these statistics, IGoR can output a list of potential recombination scenarios with their corresponding likelihoods for each sequence. As such, using the default IGoR V(D)J recombination statistics, the ten highest probability V(D)J recombination scenarios were inferred for each TCRβ-chain sequence in the training data set [11]. We then annotated each TCRβ-chain sequence with a single V(D)J recombination scenario by sampling from these ten scenarios according to the posterior probability of each scenario. We filtered these sequences for rearrangements which contained more than one trimmed nucleotide and less than fifteen trimmed nucleotides (see the "Notation" section for further details). We further subset the data to include only non-productive sequences, and used these data for all subsequent model training. After these processing and filtering steps, we used V-gene trimming length distributions from 21,193,153 non-productive sequences for all model training. To test each trained model, we used V-gene trimming length distributions from the remaining 107,121,841 productive sequences (as described in Appendix D). From this same data set, we also used J-gene trimming length distributions from 107,255,406 productive sequences and 20,204,801 non-productive sequences to test each model.

TCRα and TCRβ testing data sets: Annotated TCRα and TCRβ repertoire sequence data for 150 healthy subjects was downloaded using the link provided in the original publication [2]. In contrast to the training data cohort, this cohort contains different demographics, shallower RNA-seq based TCR-sequencing, and was processed using a different sequence annotation methods (i.e. TCRdist (version 0.0.2) [67] as described in a previous publication [2]). Sequences were split into non-productive and productive groups for model validation. From the TCRα data set, we used V-gene trimming length distributions from 123,496 non-productive sequences and 862,096 productive sequences and J-gene trimming length distributions from 141,451 non-productive sequences and 1,101,114 productive sequences to test each model. From the TCRβ data set, we used V-gene trimming length

distributions from 64,738 non-productive sequences and 1,435,153 productive sequences and J-gene trimming length distributions from 59,608 non-productive sequences and 1,496,953 productive sequences to test each model.

TCRγ testing data set: Annotated TCRγ repertoire sequence data for 23 healthy bone marrow donor subjects was downloaded from the Adaptive Biotechnologies immuneACCESS database [102]. Sequences were split into non-productive and productive groups for model validation. We used V-gene trimming length distributions from 2,403,293 non-productive sequences and 1,002,662 productive sequences and J-gene trimming length distributions from 568,824 non-productive sequences and 250,493 productive sequences to test each model.

IgH testing data sets: Annotated IgG class non-productive IgH repertoire sequence data for 9 healthy subjects was obtained from the authors of a previous publication [97]. The raw sequence data is available using the link provided in the original publication [103]. In contrast to the training data cohort, this cohort contains different demographics, shallower RNA-seq based IgH-sequencing, and was processed using a different sequence annotation method (i.e. a combination of Immcantation [104] and IgBlast [9] as described in a previous publication [97]). Further, these data are restricted to rearrangements that lead to a clonal family with at least six members.

Likewise, productive IgH repertoire sequence data for 4 healthy subjects was downloaded using the link provided in the original publication [105] and the sequences were annotated using partis (version 0.16.0) [106]. Due to the large size of this data set, 100k sequences were randomly sampled from the original data set prior to model validation. For both IgH data sets, only a single sequence from each inferred clonal family was included in each model testing data set. From these data sets, we used V-gene trimming length distributions from 160,714 non-productive sequences and 32,245 productive sequences and J-gene trimming length distributions from 297,298 non-productive sequences and 74,884 productive sequences

to test each model.

Artemis-locus SNP data set: Genome-wide SNP array data corresponding to 611 of the training data set individuals was downloaded from The database of Genotypes and Phenotypes (accession number: phs001918). Details of the SNP array data set, genotype imputation, and quality control have been described previously [57]. We only used SNP data corresponding to the Artemis-locus (rs41298872) which we previously found to be strongly associated with increasing the extent of V-gene trimming [2].

3.3.2 Notation

Let I be a set of individuals. For each subject $i \in I$, assume we have a TCR repertoire consisting of sequences indexed by k such that $k = 1, \ldots, K_i$. We assume that each sequence can be unambiguously annotated with being from a specific V-gene and J-gene sequence, and having a number of deleted nucleotides from each gene. For modeling purposes, we combine *TRB* V-gene or J-gene alleles that have identical terminal nucleotide sequences (last 24 nucleotides of each sequence) into *TRB* V-gene allele groups and *TRB* J-gene allele groups. As such, each TCR sequence is annotated with being from a V-gene allele group and J-gene allele group. Because we are requiring that each gene-allele group originates from the same *TRB* V-gene or J-gene, there may still be overlap in terms of sequence identity between allele groups. For simplicity, we orient all sequences in the 5'-to-3' direction, and use the top strand for V-gene sequences and the bottom strand for J-gene sequences. We will be introducing modeling methods as they relate to V-genes and V-gene trimming, however, with this sequence orientation, the same methods can be applied to J-genes and J-gene trimming. We will use σ to represent a gene sequence oriented in the 5'-to-3' direction and n to represent the number of nucleotides deleted from the 3' end of this sequence as we describe our modeling.

We are interested in modeling the probability of trimming n nucleotides from a given gene sequence σ, $P(n \mid \sigma)$. We can define an empirical conditional probability density function to estimate this probability. To start, we can uniformly sample from any given individual's repertoire. Let S be a random variable that represents the gene allele group sequence from such a sample. Let N be a random variable that represents the number of deleted nucleotides, which for notational convenience we assume take on a non-negative integer value (nonsensical values will have probability zero). Let $0 \leq C^{(i)}(\sigma) \leq K_i$ represent the number of TCRs that use gene allele group σ. Let $0 \leq C^{(i)}(n, \sigma) \leq K_i$ represent the number of TCRs that have gene allele group σ and n gene nucleotides deleted. With these data, we can form the empirical conditional probability density function:

$$P_{\text{emp}}(N = n \mid S = \sigma, i) = \frac{C^{(i)}(n, \sigma)}{C^{(i)}(\sigma)}. \tag{3.1}$$

Using these TCRβ repertoire data, we want to model the influence of various sequence-level parameters on $P(n \mid \sigma)$. With this assumption, let L and U be lower and upper bounds, respectively, on n such that $N' = \{L, \ldots, U\}$ is the set of all reasonable nucleotide deletion amounts. The precise location of hairpin opening and its relationship to deletion is unclear. Hence, we have chosen to define $L = 2$ since smaller trimming amounts may result from an alternative, hairpin-opening-position-related (or other) trimming mechanism. Likewise, we have chosen to define $U = 14$ since trimming amounts greater than 14 nucleotides are uncommon and could also result from an alternative trimming mechanism. We will subset the training data set, after IGoR annotation (see details in a previous section), such that we will only consider TCRs that have $2 \leq n \leq 14$. Similarly, the one existing analysis [15] exploring the relationship between sequence context and nucleotide trimming only considered TCRs that had $2 \leq n \leq 12$ for their modeling. We summarize all of the notation discussed in this section, as well as in the following sections, in Table C.1.

3.3.3 V(D)J recombination modeling assumptions

For our model, we make the following assumptions about V(D)J recombination biology:

1. During the V(D)J recombination process, the gene DNA hairpin is nicked open by a single-stranded break [26, 27, 29–31].
2. This hairpin nick occurs at the +2 position, leading to a 4 nucleotide long 3'-single-stranded-overhang (the 2 nucleotides furthest 3' are considered P-nucleotides) [29, 31]. We will discuss a sensitivity analysis to this assumption, which showed that the assumed hairpin-nick position had little impact on our model fitting, in Appendix D.
3. If any nonzero amount of the original gene sequence is deleted, all P-nucleotides will also be deleted [26, 47].
4. Nucleotide trimming occurs before N-insertion.

With these assumptions, we can resolve the nucleotide sequence on both sides of the trimming site and define mechanistically-interpretable model features using these two sequences. Specifically, we define a "trimming motif" consisting of several nucleotides on either side of the trimming site , the predicted "DNA-shape" of the nucleotides and bonds in close proximity to the trimming site, the counts of GC or AT nucleotides on either side of the trimming site beyond the "trimming motif" region (e.g. the "two-side base-count beyond"), and the sequence-independent "length" from the end of the gene to the trimming site (see Appendix D for further details). An example of how an arbitrary V-gene sequence is transformed into features for modeling is shown in Figure 3.1. We will assume that observations can be drawn from a model in which these features vary across trimming lengths n for a given gene allele group σ. We can then explore the influence of these features on the probability of trimming at a certain site given a gene sequence.

3.3.4 Defining a model covariate function

With the features summarized above, we can define a model covariate function f than contains any unique combination of parameter-specific covariate functions (Table 3.1). This function f will be the sum of each of the desired parameter-specific covariate functions. This framework allows us to generalize the existing position-weight-matrix style model [15] to a model that allows for arbitrary sequence features. For example, we replicate this PWM model using the model covariate function, $f_1(n, \sigma; \boldsymbol{\beta}^{\texttt{motif}}, a = 2, b = 4)$, where n represents the number of trimmed nucleotides, σ represents the gene allele group sequence, $\boldsymbol{\beta}^{\texttt{motif}}$ represents *motif*-specific parameter coefficients, and a and b are non-negative integer values that represent the number of nucleotides 5' and 3' of the trimming site, respectively, that are included in the "trimming motif". This function is described further in (D.2). To extend this model to a model containing *motif* parameters and *base-count-beyond* parameters, the model covariate function will be

$$f(n, \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \coloneqq f_1(n, \sigma; \boldsymbol{\beta}^{\texttt{motif}}, a, b) + f_2(n, \sigma; \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \tag{3.2}$$

where f_2 represents the *base-count-beyond* model covariate function (D.5), $\boldsymbol{\beta}^{\texttt{AT}}$ and $\boldsymbol{\beta}^{\texttt{GC}}$ represent *base-count-beyond*-specific parameter coefficients, and c represents the number of nucleotides 5' of the trimming site to be included in the base-count. We will use this *motif* and *base-count-beyond* model example to discuss the model formulation in the following sections, however, many other parameter combinations are possible. We will not define a model covariate function that contains two parameters that model the same feature. For example, *length* and *base-count-beyond* coefficients will never be included in a model covariate function together (since they both parameterize length). Likewise, *motif* and *DNA-shape* coefficients will never both be included in a model covariate function.

Table 3.1: Summary of all parameters and covariate functions for a trimming site n and gene sequence σ. Here, a and b represent the number of nucleotides 5' and 3' of the trimming site to be included in the "trimming motif", respectively, and c represents the number of nucleotides 5' of the trimming site to be included in the base-count.

Parameter	**Model coefficient variables**	**Parameter-specific covariate function**
motif parameters	$\boldsymbol{\beta}^{\texttt{motif}}$ coefficients	$f_1(n, \sigma; \boldsymbol{\beta}^{\texttt{motif}}, a, b)$ (D.2)
base-count-beyond parameters	$\boldsymbol{\beta}^{\texttt{AT}}$ and $\boldsymbol{\beta}^{\texttt{GC}}$ coefficients	$f_2(n, \sigma; \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$ (D.5)
DNA-shape parameters	$\boldsymbol{\beta}^{\texttt{shape}}$ coefficients	$f_3(n, \sigma; \boldsymbol{\beta}^{\texttt{shape}}, a, b)$ (D.7)
length parameters	$\beta^{\texttt{ldist}}$ coefficient	$f_4(n, \sigma; \beta^{\texttt{ldist}})$ (D.8)

3.3.5 Predicting trimming probabilities using conditional logistic regression

We will be using the *motif* and *base-count-beyond* parameters given by (3.2) as examples for the remainder of this section, however, we could also formulate a model with any other parameter of interest, as described in the previous section (Table 3.1). As such, we can fit a conditional logit model which posits that

$$P(n \mid \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \coloneqq \frac{\exp(f(n, \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c))}{\sum_{n' \in N'} \exp(f(n', \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c))}. \tag{3.3}$$

where N' is the set of all reasonable trimming lengths, a and b represent the number of nucleotides 5' and 3' of the trimming site to be included in the "trimming motif", respectively, c represents the number of nucleotides 5' of the trimming site to be included in the base-count parameters, and $f(n, \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c))$ is the model covariate function for the *motif* and *base-count-beyond* model given by (3.2). We will let $P(n \mid \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$ denote the conditional probability that a given gene will be trimmed by n nucleotides.

Let $y_{ik\sigma n}$ equal 1 if a gene allele group σ is trimmed by n nucleotides for TCR k from subject i, and equal 0 otherwise. With this, we can define a likelihood function, $L(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$, such that for a random sample of subjects, $L(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$, is the likelihood of the model parameters, $\boldsymbol{\beta}^{\texttt{motif}}$, $\boldsymbol{\beta}^{\texttt{AT}}$, and $\boldsymbol{\beta}^{\texttt{GC}}$, given that we observed a set of trimming amounts for a set of given genes. As such, the log-likelihood function can be

written as

$$\log L(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) = \sum_{i,k,\sigma,n} y_{ik\sigma n} \cdot \log P(n \mid \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$$

where $P(n \mid \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$ is given by (3.3). Instead of maximizing this log-likelihood directly, we may wish to aggregate the data to reduce the number of observations and simplify model fitting. Recall that for subject i, $C^{(i)}(\sigma)$ represents the number of TCRs which use gene allele group σ and $C^{(i)}(n, \sigma)$ represents the number of TCRs which have gene allele group σ and n gene nucleotides deleted. As such, $C^{(i)}(n, \sigma)$ is the count of observations which will have the same trimming probabilities $P(n \mid \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$ and will have been trimmed by n for subject i and gene allele group σ. Thus, using this aggregated data from all subjects $i \in I$, we can re-write the log-likelihood function equivalently as

$$\log L(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) = \sum_{i,\sigma,n} C^{(i)}(n, \sigma) \cdot \log P(n \mid \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c). \quad (3.4)$$

As above, for a random sample of subjects, $L(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$, is the likelihood of the model parameters, $\boldsymbol{\beta}^{\texttt{motif}}$, $\boldsymbol{\beta}^{\texttt{AT}}$, and $\boldsymbol{\beta}^{\texttt{GC}}$, given that we observed a set of trimming amounts for a set of given genes.

With this likelihood formulation, all observations in the sample get uniform treatment in the construction of the likelihood. However, subjects may differ in their repertoire size and composition for reasons other than trimming. For example, it is known that gene usage differs across subjects. Thus, to avoid having these differences pollute our $\boldsymbol{\beta}^{\hat{\texttt{motif}}}$, $\boldsymbol{\beta}^{\hat{\texttt{AT}}}$, and $\boldsymbol{\beta}^{\hat{\texttt{GC}}}$ inference, we propose a subject and gene weighting scheme.

As such, we can define the expected likelihood of a process where we first draw a subject i uniformly at random, then we sample T-cell receptor sequences from their repertoire according to a given distribution, as follows. For a single TCR sequence from such a sample,

let S be a random variable representing the gene of the sequence, and let N be a random variable representing the number of deleted nucleotides. We can sample each TCR sequence with probability $P_{\text{sample}}(N = n, S = \sigma)$ which we will specify later. Also, given random S and N, the log-likelihood of the model parameters, $\boldsymbol{\beta}^{\texttt{motif}}$, $\boldsymbol{\beta}^{\texttt{AT}}$, and $\boldsymbol{\beta}^{\texttt{GC}}$, is given by

$$\log L(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c; N, S) = \log P(N \mid S; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c).$$

With this, the expected log-likelihood of the model parameters, $\boldsymbol{\beta}^{\texttt{motif}}$, $\boldsymbol{\beta}^{\texttt{AT}}$, and $\boldsymbol{\beta}^{\texttt{GC}}$ given this random sample is given by

$$\begin{aligned}
&E[\log L(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \mid I = i] \\
&\quad = \sum_{n,\sigma} P_{\text{sample}}(N = n, S = \sigma) \cdot \log P(N = n, S = \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \\
&\quad = \sum_{n,\sigma} P_{\text{sample}}(N = n, S = \sigma) \cdot \log P(N = n \mid S = \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c).
\end{aligned}$$

We can define a new, weighted log-likelihood function, $\log L_{\text{expected}}(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$, equivalent to this expected log-likelihood:

$$\log L_{\text{expected}}(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \coloneqq \sum_i E[\log L(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \mid I = i]. \tag{3.5}$$

For a random sample of subjects, the weighted likelihood, $L_{\text{expected}}(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$, represents the likelihood of the model parameters, $\boldsymbol{\beta}^{\texttt{motif}}$, $\boldsymbol{\beta}^{\texttt{AT}}$, and $\boldsymbol{\beta}^{\texttt{GC}}$, given that we observed a set of trimming amounts for a given set of gene allele groups after weighting observations according to the sampling procedure $P_{\text{sample}}(N = n, S = \sigma)$. We can use whichever sampling procedure, $P_{\text{sample}}(N = n, S = \sigma)$, we want. For example, recall that we originally formed the empirical conditional PDFs in (3.1) for each subject i by uniformly sampling

from each TCR repertoire to get a total repertoire size of K_i:

$$P_{\text{emp}}(N = n \mid S = \sigma, i) = \frac{C^{(i)}(n, \sigma)}{C^{(i)}(\sigma)},$$

$$P_{\text{emp}}(S = \sigma \mid i) = \frac{C^{(i)}(\sigma)}{K_i},$$

and

$$P_{\text{emp}}(i) = \frac{1}{I}.$$

With this, we can define a sampling procedure equivalent to this empirical joint PDF as follows:

$$\begin{aligned} P_{\text{sample}}(N = n, S = \sigma) \coloneqq & P_{\text{emp}}(N = n, S = \sigma) \\ = & P_{\text{emp}}(n \mid \sigma, i) \cdot P_{\text{emp}}(\sigma \mid i) \cdot P_{\text{emp}}(i). \end{aligned} \tag{3.6}$$

With this sampling procedure,

$$\begin{aligned} & \log L_{\text{expected}}(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \\ & = \sum_{i,\sigma,n} P_{\text{emp}}(n \mid \sigma, i) \cdot P_{\text{emp}}(\sigma \mid i) \cdot P_{\text{emp}}(i) \cdot \log P(n \mid \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \end{aligned} \tag{3.7}$$

As such, each subject, instead of each observation, gets uniform treatment in the construction of the weighted likelihood.

While this procedure would correct for individual subjects having different repertoire sizes, it does not account for gene usage differences. To avoid having these differences pollute our $\boldsymbol{\beta}^{\hat{\texttt{motif}}}$, $\boldsymbol{\beta}^{\hat{\texttt{AT}}}$, and $\boldsymbol{\beta}^{\hat{\texttt{GC}}}$ inference, we propose a subject-independent gene allele group sampling scheme. While we could use any distribution on σ, including a uniform weight by gene allele groups, we have chosen to define:

$$P_{\text{marg}}(\sigma) = \frac{1}{I} \sum_{i} P_{\text{emp}}(\sigma \mid i).$$

We can reformulate the sampling procedure which is an empirical average per-gene-allele-group frequency such that:

$$P_{\text{sample}}(N = n, S = \sigma) \coloneqq P_{\text{emp}}(n \mid \sigma, i) \cdot P_{\text{marg}}(\sigma) \cdot P_{\text{emp}}(i). \tag{3.8}$$

With this subject-independent gene sampling procedure, we can define a weighted likelihood $L_W(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$ such that

$$\begin{aligned} &\log L_W(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \\ &\quad \coloneqq \sum_{i,\sigma,n} P_{\text{emp}}(n \mid \sigma, i) \cdot P_{\text{marg}}(\sigma) \cdot P_{\text{emp}}(i) \cdot \log P(n \mid \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \end{aligned} \tag{3.9}$$

As such, each gene and each subject get uniform treatment in the construction of the weighted likelihood.

From here, we can maximize this weighted log-likelihood, $\log L_W(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$, to estimate the log-probabilities $\boldsymbol{\beta}^{\texttt{motif}}$, $\boldsymbol{\beta}^{\texttt{AT}}$, and $\boldsymbol{\beta}^{\texttt{GC}}$, where $\boldsymbol{\beta}^{\texttt{motif}}$ is equivalent to a (log) position-weight-matrix. To estimate each coefficient, we can solve the weighted maximum likelihood estimation problem:

$$(\hat{\boldsymbol{\beta}^{\texttt{motif}}}, \hat{\boldsymbol{\beta}^{\texttt{AT}}}, \hat{\boldsymbol{\beta}^{\texttt{GC}}}) = \text{argmax}_{\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}} \ \log L_W(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \tag{3.10}$$

using the `mclogit` package in R. We can formulate a weighted maximum likelihood problem in a similar way for any model covariate function f containing a unique combination of parameter-specific covariate functions (Table 3.1).

We compare our inferred coefficients to the existing position-weight-matrix-style model which was designed and trained using least squares [15]. When replicating this model using our methods described above (i.e. the *2x4 motif* model), we note highly similar results (Figure C.1).

3.3.6 Evaluating model fit and generalizability across genes

In order to evaluate the model fit and generalizability of each model, we use a variety of training and testing data sets to train each model and calculate the log loss. We will describe our general model evaluation procedure here. We describe variations of this general model evaluation procedure in Appendix D. Let $\mathbf{T}$ represent a training data set and $\mathbf{H}$ represent a held-out testing data set. With the training set $\mathbf{T}$, we can train each model of interest as described above in (3.10). After this model fitting, we can calculate the expected per-sequence conditional log loss of the model with given coefficients, $\mathcal{M}$, for a given held-out testing set, $\mathbf{H}$, such that

$$\begin{aligned}\ell(\mathcal{M} \mid \mathbf{H}) &\coloneqq -\sum_{i,\sigma,n} P_{\text{emp}_{\mathbf{H}}}(n, \sigma, i) \cdot \log P(n \mid \sigma; \mathcal{M}) \\ &= -\sum_{i,\sigma,n} P_{\text{emp}_{\mathbf{H}}}(n \mid \sigma, i) \cdot P_{\text{emp}_{\mathbf{H}}}(\sigma \mid i) \cdot P_{\text{emp}_{\mathbf{H}}}(i) \cdot \log P(n \mid \sigma; \mathcal{M})\end{aligned} \tag{3.11}$$

where i represents a subject, n represents a trimming length, and σ represents a gene allele group. Because we are incorporating the empirically observed frequency of each subject, trimming length, and gene allele group within each "held-out testing set," $P_{\text{emp}_{\mathbf{H}}}(n, \sigma, i)$, in this formulation, the expected per-sequence conditional log loss values are guaranteed to be directly comparable between held-out testing sets with varying compositions. Models that have lower expected per-sequence conditional log loss will indicate that the model has a better fit.

3.3.7 Assessing significance of model coefficients

During model fitting, we estimated the model coefficients $\hat{\boldsymbol{\beta}^{\texttt{motif}}}$, $\hat{\boldsymbol{\beta}^{\texttt{AT}}}$, and $\hat{\boldsymbol{\beta}^{\texttt{GC}}}$ by maximizing the weighted likelihood function given by (3.9). To measure the significance of each of these model coefficients $\hat{\beta} \in \{\hat{\boldsymbol{\beta}^{\texttt{motif}}}, \hat{\boldsymbol{\beta}^{\texttt{AT}}}, \hat{\boldsymbol{\beta}^{\texttt{GC}}}\}$ we want to test whether each coefficient $\hat{\beta} = 0$. To

do this, we can first estimate the standard error of each inferred coefficient using a clustered bootstrap (with subject-gene pairs as the sampling unit). As such, for each bootstrap iterate, we sampled subject-gene pairs from the full V-gene training data set with replacement. Using this re-sampled data, we maximized the weighted likelihood function given by (3.9) to re-estimate each coefficient. We repeated this bootstrap process 1000 times and used the resulting 1000 coefficient estimates to estimate a standard error for each model coefficient. With this estimated standard error of each inferred model coefficient $\hat{\beta} \in \{\hat{\boldsymbol{\beta}^{\texttt{motif}}}, \hat{\boldsymbol{\beta}^{\texttt{AT}}}, \hat{\boldsymbol{\beta}^{\texttt{GC}}}\}$, we test whether $\hat{\beta} = 0$ by calculating the test statistic

$$T(\hat{\beta}) = \frac{\hat{\beta}}{\text{se}(\hat{\beta})} \tag{3.12}$$

and comparing $T(\hat{\beta})$ to a $N(0,1)$ distribution to obtain each P-value. We consider the significance of each model coefficient using a Bonferroni-corrected threshold. To establish the threshold, we corrected for the total number of model coefficients being evaluated in the given model.

3.3.8 Evaluating model coefficient variation in the context of SNPs

With the *motif* and *base-count-beyond* model, we are interested in quantifying variation in model coefficients in the context of genetic variations within the gene encoding the Artemis protein that were previously-identified as being associated with increasing the extent of trimming [2]. Recall that we trained this model using the model covariate function given by (3.2). During model fitting, we estimated the model coefficients $\hat{\boldsymbol{\beta}^{\texttt{motif}}}$, $\hat{\boldsymbol{\beta}^{\texttt{AT}}}$, and $\hat{\boldsymbol{\beta}^{\texttt{GC}}}$ by maximizing the weighted likelihood function given by (3.9).

We have previously identified a set X of single-nucleotide-polymorphisms (SNPs) within the gene encoding the Artemis protein that are significantly associated with increasing the extent of trimming [2]. For each SNP $x \in X$ and individual $i \in \{1, \ldots, I\}$, we measure

the number of minor alleles in the genotype, $g_{ix} \in \{0, 1, 2\}$. We are interested in whether each of the inferred model coefficients $\hat{\beta} \in \{\hat{\boldsymbol{\beta}^{\texttt{motif}}}, \hat{\boldsymbol{\beta}^{\texttt{AT}}}, \hat{\boldsymbol{\beta}^{\texttt{GC}}}\}$ vary in the context of genotype for each genetic variant $x \in X$. As such, for each SNP of interest, we can adapt the *1x2 motif + two-side base-count beyond* model covariate function to allow for genotype-specific variation of each model coefficient by incorporating additional interaction coefficients $\beta_x \in \{\boldsymbol{\beta}_x^{\texttt{motif}}, \boldsymbol{\beta}_x^{\texttt{AT}}, \boldsymbol{\beta}_x^{\texttt{GC}}\}$ to model the relationship between each model parameter and the SNP x genotype. We can then estimate the coefficients of this new model, $\hat{\boldsymbol{\beta}^{\texttt{motif}}}$, $\hat{\boldsymbol{\beta}^{\texttt{AT}}}$, $\hat{\boldsymbol{\beta}^{\texttt{GC}}}$, $\hat{\boldsymbol{\beta}^{\texttt{motif}}}_x$, $\hat{\boldsymbol{\beta}^{\texttt{AT}}}_x$, and $\hat{\boldsymbol{\beta}^{\texttt{GC}}}_x$, as before by maximizing the weighted likelihood given by (3.9) using the adapted model covariate function. We can measure the significance of each of the model coefficients using the methods described in the previous section. Ultimately if a SNP-coefficient interaction term $\hat{\beta}_x \in \{\hat{\boldsymbol{\beta}^{\texttt{motif}}}_x, \hat{\boldsymbol{\beta}^{\texttt{AT}}}_x, \hat{\boldsymbol{\beta}^{\texttt{GC}}}_x\}$ is significant, we can conclude that the corresponding model coefficient $\hat{\beta}$ varies significantly in the context of the genotype of SNP $x \in X$. We use this same procedure to evaluate whether each model coefficient varies in the context of each SNP of interest.

3.3.9 Code availability

R code implementing the modeling described here is available at `https://github.com/magdalenarussell/mechanistic-trimming`.

Chapter 4

STATISTICAL ANALYSIS OF REPERTOIRE DATA DEMONSTRATES THE INFLUENCE OF MICROHOMOLOGY IN V(D)J RECOMBINATION

V(D)J recombination is an essential process for generating diverse B cell receptors (BCRs) and T cell receptors (TCRs). In this process, single V-, D- (if present), and J-genes are randomly selected from a pool of germline gene segments, then edited and joined together to form a uniquely recombined receptor sequence. Previous *in vitro* experiments have suggested that short stretches of sequence homology between gene ends, known as microhomology, can play a significant role in the V(D)J recombination process [26, 44, 45, 45, 46, 107–111]. This raises the question of whether microhomology impacts V(D)J recombination *in vivo*, particularly in terms of recombination outcomes in humans with intact recombination machinery. Understanding this has practical implications for V(D)J recombination sequence *annotation*. Annotation means inferring the specific V(D)J recombination editing and joining processes that produced each sequence, forming the basis for many downstream B cell and T cell repertoire analyses. In this chapter, we use statistical inference on high-throughput human TCR repertoire data to assess how microhomology influences various steps of the V(D)J recombination process.

In order to more fully set the stage, we will now summarize the relevant biological context. V(D)J recombination begins when the recombination activating gene (RAG) protein

complex aligns two randomly chosen genes, removes the intervening chromosomal DNA between the two genes, and forms a hairpin loop at the end of each gene [21–23]. Each hairpin loop is then nicked open by the Artemis:DNA-PKcs complex [23, 29, 31]. Hairpin opening most frequently occurs at position +2, where position 0 refers to the edge of the hairpin and position -1 refers to the last nucleotide on the 5' strand [29], however, other hairpin opening positions are also possible [29, 31]. The Ku heterodimer (Ku70/Ku80) can bind to each nicked gene end and recruit non-homologous end joining factors, in any order, to repair the double stranded break [44, 109, 112]. From here, it is likely that the processing of the two gene ends occurs iteratively, with multiple rounds of action by a nuclease, polymerase, and ligase which eventually leads to a joining event to combine the two gene fragments [109, 111].

The various possible processing steps involved in this iterative end-joining stage are as follows. Nucleotides can be trimmed from each gene end through a mechanism suggested to involve the Artemis nuclease [2, 25, 27, 30, 32–34, 45, 80, 81]. Nucleotide deletion is thought to occur in a sequence-dependent fashion; for example, sequences with high AT content have been found to experience greater nucleotide loss than those with high GC content [25–27, 30], and the extent of deletions has been shown to depend on local nucleotide identity [3, 15], as well as sequence breathing capacity and length [3]. Additionally, non-template-encoded nucleotides, known as N-insertions, can be added by terminal deoxynucleotidyl transferase (TdT) [35–37]. TdT has a bias for the addition of purine-purine and pyrimidine-pyrimidine di-nucleotides suggesting that nucleotide addition depends on the previous addition [15, 26]. Further, nucleotide addition lengths and composition have been shown to depend on the presence (or absence) of nucleotide trimming at the gene ends [113]. Joining of the two gene ends is then carried out by XRCC4:DNA ligase IV, a flexible ligase that can ligate across gaps and incompatibilities between the ends, along with additional end-joining factors like XLF and PAXX that stabilize the ends, and polymerases that fill in gaps [46, 114–117].

The presence of microhomology, while not required, has been suggested to bias the out-

come of the random V(D)J recombination processing steps. Microhomology can occur in several forms: (1) **terminal microhomology**[1], found at the ends of genes prior to trimming/insertion and encoded in the germline; (2) **interior microhomology**, located *within* the sequences and also germline-encoded; and (3) **insertion-dependent microhomology**, created by N-insertions and not encoded in the germline. Because terminal and interior microhomology are both germline-encoded, we will collectively refer to them as *germline-encoded microhomology*. If present, terminal microhomology can directly guide ligation without additional processing. In contrast, interior and insertion-dependent microhomology may necessitate deletions or further N-insertions before microhomology-mediated ligation can occur. This chapter will focus exclusively on germline-encoded microhomologies, excluding insertion-dependent microhomology.

Experiments *in vitro* and with model organisms have suggested that microhomology (i.e. 1-4 nucleotides) is an important factor in V(D)J recombination. Although microhomology between gene ends is not essential for joining (Figure 4.1A part (i)) [42, 43], it has been shown to improve joining efficiency and bias the outcome towards using the microhomologous region to guide trimming and ligation (Figure 4.1A part (ii)) [26, 44–46, 109–111]. For example, reconstitution experiments suggest that sequences with microhomology can stabilize gene ends without requiring additional end-joining factors like XLF and PAXX and germline-encoded microhomology may reduce the necessity for template-independent addition by polymerase-μ and TdT [46], possibly explaining the enhanced ligation efficiency. *In vitro* studies show that 1 or 2 nucleotides of germline-encoded microhomology are present in nearly 60% of ligated coding joints in the absence of TdT [26], with similar observations reported in neonatal mice when TdT levels are low [107, 108]. However, this frequency drops substantially when TdT

[1]In this work, we use the term "terminal microhomology" to describe *germline-encoded* microhomologous nucleotides located at gene ends (prior to trimming). However, other sources [45, 109] often use the term more broadly to describe all microhomologous nucleotides located at gene ends, including both germline-encoded nucleotides and those generated through N-insertion.

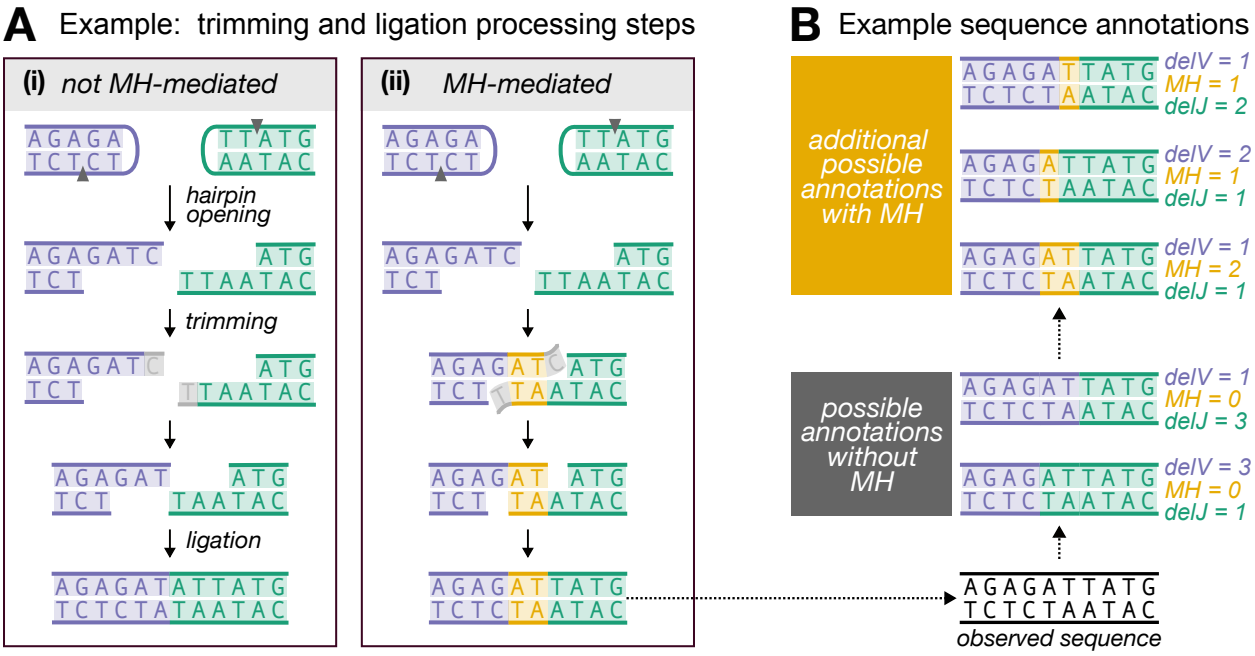

Figure 4.1: **(A)** Illustration of how germline-encoded microhomology (MH) could affect trimming and/or ligation during V(D)J recombination. We use sequences without N-insertions to quantify these effects, leveraging germline V- and J-gene sequences to identify potential MH-mediated ligation events. The example shows germline-encoded interior microhomologous regions (yellow) and trimmed nucleotides (gray) for a V-gene (purple) and J-gene (green), highlighting two regimes: (i) no MH influence and (ii) MH-mediated trimming and ligation. In these scenarios, germline-encoded microhomologous regions are classified as terminal microhomology when they facilitate the ligation of untrimmed sequences, and as interior microhomology when they facilitate the ligation of trimmed sequences. MH could affect trimming, ligation, or both, leading to distinct sequences regardless of whether the genes are trimmed equally or differently. We illustrate how other forms of microhomology could affect trimming and ligation during V(D)J recombination within Figure E.1. **(B)** Illustration of possible V(D)J recombination annotations for sequences lacking N-insertions that potentially ligate with MH. Existing annotation software does not account for MH and assigns shared nucleotides to only one sequence (gray box annotations). However, additional annotations that incorporate MH (yellow box) are possible but are not considered by existing software. Example trimming scenarios, given by delV and delJ, and ligation scenarios, given by MH, for each possible annotation are shown. Our modeling aims to consider all annotations, both with and without MH, for each sequence.

is present, as TdT-mediated additions are thought to create stronger insertion-dependent microhomology [26, 45, 109]. The involvement of microhomology in ligation appears to be more complex when it is not present at sequence ends or generated through nucleotide

addition. Most gene ends lack terminal microhomology after hairpin opening but share interior microhomology [31, 46, 110]. In such cases, the Artemis–DNA-PKcs complex has been shown to trim gene ends to expose interior microhomology (Figure 4.1A part(ii)) [46, 110]. Figure E.1 provides an extended overview illustrating how different forms of microhomology could influence V(D)J recombination.

This essential biochemical work has demonstrated that microhomology can significantly affect V(D)J recombination, however, it does not demonstrate its importance for shaping V(D)J recombination in humans. In addition to being an issue of intrinsic interest, the role of microhomology has practical implications as well: if microhomology impacts the probability of V(D)J recombination annotations (i.e. numerical histories of recombination events such as gene choice, trimming, insertion, ligation, etc.), then corresponding terms should be incorporated into software that infers recombination probabilities. This would ensure that additional annotations involving microhomology are also considered (Figure 4.1B).

Statistical inference on high-throughput repertoire sequencing datasets allows exploration of the *in vivo* V(D)J recombination mechanism in humans. In fact, existing probabilistic models of V(D)J recombination, such as IGoR [11], have provided interesting and important insights about the natural underlying mechanism by learning statistics of V(D)J recombination. These models have revealed significant dependencies between recombination events, such as gene usage and trimming, and have provided estimates of the overall probabilities of generating specific TCR sequences, thereby helping to disambiguate the effects of generation from selection [11, 15]. Similar statistical approaches have been successfully applied to understand the sequence-dependent process of nucleotide trimming, revealing significant connections between trimming patterns and local sequence identity, length, and wider GC content [3]. However, to our knowledge, no probabilistic models of V(D)J recombination incorporating microhomology have been developed.

In this chapter, we explore the extent to which germline-encoded microhomology bi-

ases trimming and ligation during V(D)J recombination using statistical inference on high-throughput TCRα repertoire sequencing data [118, 119]. We have designed a flexible probabilistic modeling framework, allowing us to quantify the extent to which germline-encoded microhomology biases trimming and ligation probabilities. Our results show that the presence of germline-encoded microhomology significantly increases trimming and ligation probabilities, and is an important predictor of the choices made in these processes. These observations are consistent with sequences from an independent TCRα validation dataset, as well as with sequences from other receptor loci such as TCRγ. Additionally, we demonstrate that explicitly including microhomology-related terms in our model substantially impacts sequence annotation probabilities and overall V(D)J recombination annotation rankings. Together, these findings enhance our understanding of the involvement of germline-encoded microhomology in the V(D)J recombination process and highlight the importance of accounting for microhomology-related effects in receptor sequence processing and analysis.

4.1 Materials and Methods

4.1.1 Terminology

In this chapter, we investigate the mechanisms of trimming and ligation as they occur between V- and J-gene pairs during V(D)J recombination. We will use these terms throughout the chapter:

- **Trimming scenario**: A specific pair of trimming events, one at the V-gene end and one at the J-gene end.
- **Ligation scenario**: A specific number of germline-encoded microhomologous nucleotides shared between the trimmed V-gene and J-gene, facilitating their ligation. The possible ligation scenarios for a given V-J gene pair are determined by their germline

sequences and the extent of trimming.

- **Joint trimming and ligation scenario probability**: The normalized probability of a particular combination of trimming and ligation scenarios occurring for a V-J gene pair, considering all possible trimming-ligation combinations for that pair.
- **V(D)J recombination annotation**: A specific set of V(D)J recombination events that produce a sequence, including trimming, insertion, and ligation scenarios.
- **V(D)J recombination annotation probability**: The normalized probability of a particular V(D)J recombination annotation for an observed sequence, calculated from all possible annotations for that sequence. We restrict our analysis to sequences without N-insertions such that we can derive these probabilities from joint trimming and ligation scenario probabilities and normalize over all possible scenario combinations for that specific observed sequence.

4.1.2 Data and data processing overview

To explore trimming and ligation patterns, we analyzed TCRα-immunosequencing data from 10 individuals [118, 119]. The *TRA* locus was chosen for its higher sequence diversity between joining genes (V- and J-gene pairs) compared to the *TRB* locus.

We used the IGoR software (version 1.4.0), designed to learn unbiased recombination statistics from immune sequence reads [11], to infer possible V(D)J recombination annotations and their associated likelihoods for each sequence. Each annotation consists of inferred V- and J-gene assignments, trimming lengths, and the number of N-insertions. For each sequence, we processed these annotations in two steps. First, we sampled a single V- and J-gene assignment and N-insertion amount based on their posterior probabilities. Sequences with N-insertions were excluded to focus on germline-microhomology-mediated ligation events, as

N-insertions complicate ligation pattern analysis due to their unknown nucleotide composition prior to ligation and indicate that *germline*-microhomology-mediated ligation did not occur. In these training data, we found that roughly 5% of sequences contained zero inferred N-insertions.

Next, given the IGoR-inferred V- and J-gene assignments and N-insertion amounts, we determined the set of possible trimming and ligation scenarios for each sequence. Since IGoR does not account for microhomology and assigns shared nucleotides to only one sequence, we did not use the corresponding IGoR-inferred trimming annotation. Instead, we adapted this IGoR-inferred trimming annotation to account for germline-encoded microhomology. This approach allowed us to generate a set of possible trimming and ligation scenario annotations for each sequence, including those that involve germline-encoded microhomologous nucleotides (see Figure 4.1B and Appendix F for details).

Additionally, TCR sequences can be categorized as "productive" if they code for a functional protein, or "non-productive" otherwise, arising from out-of-frame recombination or presence of stop codons. Each T cell can undergo recombination at two alleles; if the first is non-productive and the second successful, both sequences can be sequenced as part of the repertoire. Non-productive sequences do not generate proteins for thymic selection, and their recombination statistics should reflect only the V(D)J recombination process [15, 19, 20]. In contrast, productive sequence statistics reflect both recombination and selection. To study nucleotide trimming and ligation during V(D)J recombination without selection effects, we included only non-productive sequences in our training dataset. In these data, we found that roughly 67% of sequences were non-productive.

To validate our findings, we also analyzed productive sequences from the training dataset and both productive and nonproductive sequences from independent TCRα-immunosequencing data from 10 healthy individuals [119] and TCRγ-immunosequencing data from 23 healthy bone marrow donors [102]. These validation datasets underwent the same IGoR-based an-

notation and filtering procedures as used for the training dataset.

Further details on these datasets and processing steps are provided in Appendix F.

4.1.3 Modeling assumptions

We explore the impact of germline-encoded microhomology on V(D)J recombination by modeling the joint probability of trimming and ligation scenarios given V-gene and J-gene sequences. Our approach relies on the following biological assumptions:

1. Nucleotide trimming precedes ligation [120]
2. Each gene's DNA hairpin is opened by a single-stranded break during the early stages of V(D)J recombination [26, 27, 29–31].
3. This hairpin nick typically occurs at the +2 position, producing a 4-nucleotide 3'-overhang with two 3'-most nucleotides being considered P-nucleotides [29, 31]
4. If any part of the original gene sequence is deleted, all P-nucleotides will also be deleted [26, 47].

To simplify our analysis, we consider only the "top" strand for V-genes (5'-to-3') and the "bottom" strand for J-genes (3'-to-5'), consistent with the most common overhang polarities. Trimming is indexed from the 3' end of each strand, with trimming sites corresponding to specific coding sequence positions. Figure E.2 illustrates this sequence orientation along with the corresponding definitions.

4.1.4 Notation and modeling set-up

In order to set up our model, we will now summarize relevant notation. We uniformly sample a sequence, X, from a TCRα repertoire of filtered sequences. The following variables are random due to the choice of X, but are deterministic given X, as they are determined

by sampling from the recombination annotations inferred by IGoR based on their posterior probabilities. Let V and J be random variables representing the V-gene and J-gene, respectively, and I be a random variable representing the number of N-insertions. Let Q represent the productivity of the observed sequence, which can be either productive or non-productive. We define VJ as an ordered pair of IGoR-inferred genes: $\mathrm{VJ} = (\mathrm{V}, \mathrm{J})$. Let MH be a random variable denoting the count of shared germline-encoded microhomologous nucleotides in the ligated sequence, and delV and delJ be random variables representing the number of nucleotides deleted from the V- and J-gene, respectively. Together, we define $\mathrm{delVJ} = (\mathrm{delV}, \mathrm{delJ})$ as the pair of trimming lengths (a "trimming scenario") and M as a "ligation scenario."

For notational convenience we assume delV and delJ each take on an integer value on the interval $[-2, \ldots, 14]$, where values outside this range are considered nonsensical and assigned a probability of zero. Negative values indicate P-nucleotide deletions: a deletion of 0 means the deletion stops at the end of the germline gene sequence (e.g. two P-nucleotides are trimmed off), while a deletion of -2 indicates no deletion of P-nucleotides or gene sequence nucleotides. This indexing is consistent with the IGoR software [11] and illustrated in Figure E.2B.

Existing annotation tools like IGoR [11] do not account for microhomology and attribute shared nucleotides to only one sequence when inferring trimming scenario annotations. Instead of using IGoR-inferred trimming annotations directly, we construct a set of possible trimming and ligation scenarios for each sequence, including those involving germline-encoded microhomologous nucleotides, based on the observed sequence and known germline gene sequences. As such, given a sequence X with gene pair VJ and zero N-insertions ($\mathrm{I} = 0$), the set of possible trimming and ligation scenario annotations is described by combinations of delVJ and MH (as illustrated in Figure 4.1B). While both delVJ and MH can be considered random variables, they are dependent on one another—meaning the possible values

of MH are constrained by delVJ and vice versa. The resulting set, A_X, includes all feasible trimming and ligation scenarios consistent with X. This set is deterministic given X, but random due to the sampling of X. Details of the procedure to construct A_X are provided in Appendix F.

Our goal is to model trimming and ligation scenario probabilities given V- and J-gene pairs. To estimate the empirical conditional probability density function, let $C(\mathrm{VJ}, \mathrm{Q}, \mathrm{I} = 0)$ represent the count of TCRs within a sampled repertoire with productivity Q, using gene pair VJ, and with zero N-insertions. Let $C(\mathrm{delVJ}, \mathrm{MH}, \mathrm{VJ}, \mathrm{Q}, \mathrm{I} = 0)$ represent the count with trimming scenario delVJ, ligation scenario MH, zero N-insertions, productivity Q, and gene pair VJ. The empirical conditional probability density function is defined as:

$$P_{\mathrm{emp}}(\mathrm{delVJ}, \mathrm{MH} \mid \mathrm{VJ}, \mathrm{Q}, \mathrm{I} = 0) = \frac{C(\mathrm{delVJ}, \mathrm{MH}, \mathrm{VJ}, \mathrm{Q}, \mathrm{I} = 0)}{C(\mathrm{VJ}, \mathrm{Q}, \mathrm{I} = 0)}.$$

To achieve our goal, we will train a conditional logit model, a type of logistic model designed to model discrete choices among multiple alternatives. Specifically, we aim to model $P(\mathrm{delVJ}, \mathrm{MH} \mid \mathrm{VJ}, \mathrm{Q}, \mathrm{I} = 0)$ using sequence-level parameters, including those that capture germline-microhomology-related effects, with our TCRα repertoire training dataset. However, because the true trimming and ligation annotations $(\mathrm{delVJ}, \mathrm{MH})$ for each sequence are latent, we cannot directly compute this probability density. To address this challenge, we assign probabilities to each potential annotation based on model likelihoods. Since these probabilities depend on model parameters, we use an expectation-maximization algorithm for parameter inference, which we describe in detail in subsequent sections. We summarize all the notation discussed in this section, as well as in the following sections, in Table C.1.

4.1.5 Model formulation

In our previous work, we established that local nucleotide identities at trimming sites (the "trimming motif") and the counts of GC or AT nucleotides beyond these motifs (the "3' base count" and "5' base count") are strong predictors of trimming probabilities for single gene sequences [3]. Building on this foundation, we have integrated these established model features with newly developed germline-microhomology-related features to assess their combined effects on trimming and ligation processes (see Figure E.2 and Appendix F for detailed definitions).

To this end, we developed a two-step conditional logit model to evaluate the joint probabilities of trimming and ligation scenarios for V- and J-gene pairs. The model describes a generative process in two steps:

1. **Trimming scenario choice**: The probability $P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} = 0)$, of choosing a trimming scenario delVJ for a given V-J gene pair VJ, sequence productivity Q, and N-insertion amount I = 0. This choice is determined by the established "trimming motif", "3' base count", and "5' base count" parameters for each gene, in addition to a new parameter that quantifies the effect of germline-encoded microhomology on trimming. Specifically, this parameter measures the importance of the average number of germline-encoded microhomologous nucleotides between two trimmed sequences, a value that varies depending on the chosen trimming scenario. We denote the set of trimming-related parameters by $\boldsymbol{\beta}_{\texttt{trim}}$.

2. **Ligation scenario choice**: The probability, $P(\text{MH} \mid \text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0)$, of choosing a ligation scenario MH for a given trimming scenario delVJ, V-J gene pair VJ, sequence productivity Q, and N-insertion amount I = 0. This choice is determined by a novel microhomology parameter related to ligation, which quantifies the importance of the number of germline-encoded microhomologous nucleotides that ultimately appear in

the final ligated sequence. We denote this set of ligation-related parameters by $\boldsymbol{\beta}_{\texttt{lig}}$.

All of these modeling parameters are summarized in Table E.2, illustrated in Figure E.2D, and described in detail in Appendix F.

The mental model of this two-step process is that trimming occurs first, independently of ligation, and then ligation occurs, conditioned on the trimming scenario. However, $P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} = 0)$ will be parameterized by both trimming- and ligation-related parameters ($\boldsymbol{\beta}_{\texttt{trim}}$ and $\boldsymbol{\beta}_{\texttt{lig}}$) because the model is conditioned on sequence productivity (Q), which is jointly determined by trimming and ligation. This dependency ensures that trimming probabilities properly account for how productivity constraints prune the space of possible ligation scenarios associated with each trimming scenario, correcting for any biases introduced by this non-uniform pruning (see Appendix F for more details). Despite this dependency, the trimming-related parameters ($\boldsymbol{\beta}_{\texttt{trim}}$) and ligation-related parameters ($\boldsymbol{\beta}_{\texttt{lig}}$) are still designed to capture the distinct effects of various sequence-level features on trimming and ligation, respectively.

The joint probability of selecting a trimming scenario delVJ and a ligation scenario MH for a V-J gene pair VJ, sequence productivity Q, and zero N-insertions can thus be factored and modeled using trimming and ligation parameters $\boldsymbol{\beta}_{\texttt{trim}}$ and $\boldsymbol{\beta}_{\texttt{lig}}$ as:

$$P(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}}) :=$$
$$P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}}) \times P(\text{MH} \mid \text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{lig}}).$$

Figure 4.2 illustrates the two-step structure of this model and the decision-making process for an example V-J gene pair.

These parameters, $\boldsymbol{\beta}_{\texttt{trim}}$ and $\boldsymbol{\beta}_{\texttt{lig}}$, are designed to quantify how sequence-level features, particularly germline-encoded microhomology, influence trimming and ligation choices during V(D)J recombination. Importantly, the magnitude of microhomology's influence in

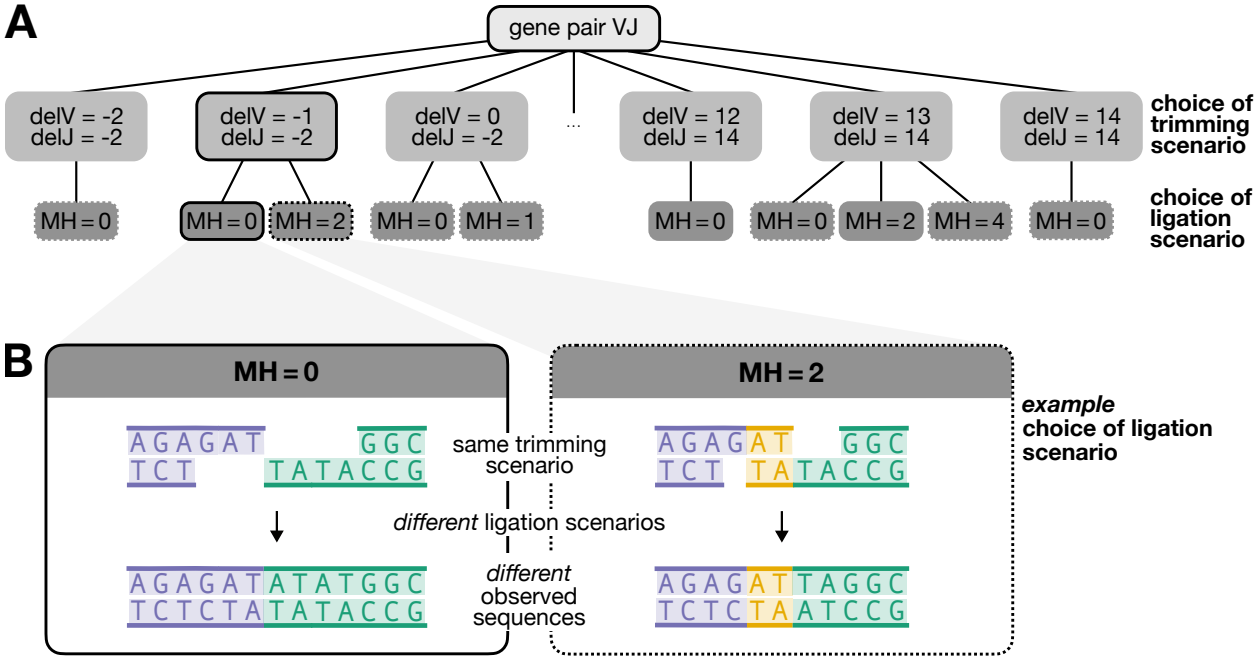

Figure 4.2: (A) Schematic of trimming and ligation choices for an arbitrary V-J gene pair, denoted by the random variable VJ. The first choice is the trimming scenario, represented by the random variable delVJ, which consists of a pair of V- and J-gene trimming amounts delV and delJ (e.g. each can range from -2 to 14 nucleotides). The next choice is the ligation scenario, represented by the random variable MH, which captures the number of germline-encoded microhomologous nucleotides used. The available ligation scenario choices depend on the germline sequences of the two genes being joined. Trimming and ligation scenarios resulting in productive and nonproductive sequences are shown in solid and dashed boxes, respectively. **(B)** Illustration of the possible ligation scenarios for an example pair of trimmed sequences. The chosen ligation scenario affects the resulting observed sequence. The trimmed V-gene sequence is shown in purple, the trimmed J-gene sequence in green, and germline-encoded interior microhomologous nucleotides in yellow. Deletions are indexed such that a deletion of 0 corresponds to the end of the germline gene sequence (two P-nucleotides trimmed) and -2 corresponds to the full sequence (no P-nucleotides trimmed), as illustrated in Figure E.2B. The ligation choices represented by $\mathrm{MH} = 0$ and $\mathrm{MH} = 2$ correspond to scenarios where zero or two germline-encoded microhomologous nucleotides (shown in yellow) are used to ligate the sequences, as reflected in the final observed sequence. Germline-encoded microhomologous regions are classified as terminal microhomology when they facilitate the ligation of untrimmed sequences, and as interior microhomology when they facilitate the ligation of trimmed sequences.

guiding these choices is quantified by these conditional logit model parameters, highlighting its role in the recombination process. We validated the model's ability to detect these effects through a series of simulations (see Appendix F). In order to assess the significance of germline-microhomology-related terms in downstream analyses, such as in V(D)J recombination sequence annotation, we designed the model with the flexibility to include or exclude germline-microhomology-related parameters for both trimming and ligation decisions.

4.1.6 Model training

We trained our conditional logit model using non-productive sequences without N-insertions and their corresponding sets of possible trimming and ligation scenarios (as described earlier). Training this model is complex because the true trimming and ligation scenarios for each sampled sequence are latent variables that depend on the model parameters. To estimate probabilities for each potential scenario, we assigned likelihoods based on our model and used an expectation-maximization (EM) algorithm for parameter inference.

Standard regression methods in R or Python could not support this type of optimization, so we implemented the EM algorithm using the `JAX` and `JAXopt` packages in Python, which support automatic differentiation [121, 122]. This algorithm converged within twenty-five iterations (Figure E.3). Further details about the EM algorithm and model formulation are provided in Appendix F.

4.1.7 Assessing significance of model parameters

When training our model, we infer a set of model parameters $\hat{\boldsymbol{\beta}} = \{\hat{\boldsymbol{\beta}}_{\texttt{trim}}, \hat{\boldsymbol{\beta}}_{\texttt{lig}}\}$ where $\boldsymbol{\beta}_{\texttt{trim}}$ are trimming-related parameters and $\boldsymbol{\beta}_{\texttt{lig}}$ are ligation-related parameters. Since the model is a conditional logit model, each parameter represents the change in the $\log_{10}$ odds of trimming and/or ligating at a specific scenario for a unit increase in the corresponding

feature value, while holding all other features constant. To assess the significance of each individual parameter $\hat{\beta} \in \hat{\boldsymbol{\beta}}$, we test the null hypothesis that $\hat{\beta} = 0$. This approach enables us to understand the contribution of each parameter separately, allowing us to evaluate the impact of specific sequence features, such as the extent of germline-encoded microhomology, on the probability of recombination events.

To test significance, we estimate the standard error of each inferred parameter using a bootstrap method, with observed sequences as the sampling unit. For each bootstrap iteration, we sample sequences from the training dataset with replacement and train a new model to re-estimate the parameters. This process is repeated 1000 times, resulting in 1000 parameter estimates, which we use to calculate the standard error for each parameter. Using these standard errors, we calculate the test statistic:

$$T(\hat{\beta}) = \frac{\hat{\beta}}{\text{se}(\hat{\beta})}.$$

We compare $T(\hat{\beta})$ to a $N(0, 1)$ distribution to obtain each p-value. We assess the significance of each model parameter using a Bonferroni-corrected threshold, adjusting for the total number of parameters being evaluated in the model.

4.1.8 Validating model using likelihood ratio testing

To determine whether adding the germline-microhomology-related terms significantly improves our model's fit to the observed data, we use a likelihood ratio test (LRT) to compare our full model that includes these terms to a simpler model that excludes them. This approach enables us to assess the collective impact of adding a set of parameters—in this case, the germline-microhomology-related parameters—to the model.

The LRT statistic compares the log-likelihoods of the two nested models:

$$\mathrm{LR} = 2 \times (\mathcal{L}_{\texttt{MH}} - \mathcal{L}_{\texttt{noMH}}).$$

Here, $\mathcal{L}_{\texttt{MH}}$ is the log-likelihood for the model with microhomology terms (defined in Appendix F, (F.20)), while $\mathcal{L}_{\texttt{noMH}}$ is the log-likelihood for the simpler model without these terms. The LRT statistic approximately follows a chi-square distribution with degrees of freedom equal to the number of additional parameters in the more complex model (e.g. two germline-microhomology-related parameters in this case).

This test allows us to calculate a p-value for the likelihood ratio, which indicates whether the inclusion of germline-microhomology-related parameters significantly improves model fit. The LRT is particularly useful for evaluating the collective contribution of related parameters, rather than individual effects. While we use bootstrap testing to assess the significance of individual parameters (as described in the previous section), the LRT enables us to evaluate the combined impact of adding germline-microhomology-related terms, allowing us to determine whether these terms are collectively biologically meaningful in the context of the observed data.

4.2 Data and code availability

Code implementing the modeling described is available using the following links: `https://github.com/magdalenarussell/microhomology` and `https://doi.org/10.6084/m9.figshare.27737685`. The data used in this study were previously published and can be accessed through the Adaptive Biotechnologies immuneACCESS database via the links provided in the original publications [102, 118, 119].

4.3 Results

4.3.1 Germline-encoded microhomology significantly increases probabilities of both trimming and ligation events

Complementary sequence regions capable of forming microhomologous regions during V(D)J recombination are common between germline V- and J-genes in the *TRA* locus. The median average number of germline-encoded microhomologous nucleotides across the ensemble of possible trimming scenarios for these germline V- and J-gene pairs is 0.1978 (Figure E.4). This median corresponds to 1.3149 possible ligation scenarios per trimming scenario (Figure E.5 and Figure E.6). Given that a median of exactly one ligation scenario per trimming scenario would indicate all V(D)J recombination annotations involve zero germline-encoded microhomology, this suggests that many trimming scenarios allow for multiple ligation outcomes, both with and without germline-encoded microhomology. Additionally, complementary sequence regions and their corresponding ligation scenario options are distributed across trimming scenarios depending on the specific V- and J-gene pair (Figure E.4 and Figure E.6). This distribution highlights the potential for both interior and terminal microhomology to influence trimming and ligation outcomes.

To quantify the effects of germline-encoded microhomologous nucleotides on trimming and ligation, we employed our model, which incorporates various sequence-level parameters, including those related to germline-encoded microhomology. We validated the model's capability to detect germline-encoded microhomology effects through a series of simulations, designed to generate sequences by sampling trimming and ligation scenarios under different microhomology regimes: no germline-encoded microhomology effect, germline-encoded microhomology affecting either trimming or ligation choices exclusively, and germline-encoded microhomology influencing both. After training our model using each of these simulated datasets, we confirmed its sensitivity to detecting variable germline-encoded microhomology

effects across different conditions (Figure E.7).

We then fit our model to the real TCRα training dataset to quantify the actual effects of germline-encoded microhomology, along with other sequence-level features, on the probabilities of trimming and ligation events. Since the model is a conditional logit model, each model parameter reflects the change in the $\log_{10}$ odds of trimming and/or ligating at a specific scenario due to an increase in the corresponding feature value, assuming all other features remain constant. We assessed the significance of each model parameter's influence on trimming and ligation event probabilities by estimating their standard errors with bootstrap methods and applying a z-test to obtain a p-value (see Methods). We used a Bonferroni-corrected significance threshold of 0.0016, adjusted for the total number of model parameters, and report parameters on the $\log_{10}$ scale. Our results indicate that the number of germline-encoded microhomologous nucleotides between two sequences substantially influences both trimming (parameter = 0.4484) and ligation (parameter = 0.1272) outcomes, with both effects being highly significant (p-values smaller than machine tolerance, $p \simeq 0$) (Figure 4.3A).

This relationship is further demonstrated by notable increases in joint trimming and ligation probabilities for scenarios with more germline-encoded microhomology, as illustrated in Figure 4.3B, which highlights two trimming and ligation scenarios from the most common V-J gene pair, TRAV41*01 and TRAJ45*01. While the influence of germline-encoded microhomology on trimming was stronger than on ligation, these effects appear to be interdependent. Interestingly, when training the model using sequences containing N-insertions (indicating a lack of ligation solely dependent on germline-encoded microhomology), germline-encoded microhomology had a small but significant effect on trimming probabilities (parameter = 0.0059; p-value smaller than machine tolerance, $p \simeq 0$) (Figure E.8). This demonstrates that germline-encoded microhomology may independently influence trimming, suggesting a nuanced role beyond its interaction with ligation. However,

it is possible that sequences containing N-insertions were ligated using microhomologous nucleotides derived from both N-insertions and germline-encoded regions, which could contribute to the observed trimming-related effects and complicate the interpretation of these signals.

Returning to the original model, in addition to germline-encoded microhomology effects, we identified significant "trimming motif", "3' base count", and "5' base count" parameters for the probabilities of both V- and J-gene trimming events. These parameters, previously introduced in our analyses of trimming patterns for single V- and J-gene sequences [3], showed results consistent with our previous work. As in our prior work, the local sequence context ("trimming motif") for each gene was modeled using a position weight matrix from a three-nucleotide window around each trimming site. We observed similar patterns for both V-gene and J-gene local trimming contexts, where C and A nucleotides had the largest influence on trimming outcomes (Figure 4.3A). The "5' base count" and "3' base count" parameters reflect how upstream and downstream AT and GC nucleotide composition influence trimming probabilities. These features are based on the raw counts of AT and GC nucleotides 5' and 3' of the trimming motif. The 5' base count parameters act as a proxy for sequence-breathing effects, indicating a preference for GC content upstream of the motif. In contrast, the 3' base count parameters capture two effects: the absolute position of the trimming site, as the total AT and GC counts downstream correspond to this position, and sequence-breathing effects driven by AT and GC content downstream of the motif. Our analysis showed that increasing GC nucleotides 5' of the motif (which decreases sequence-breathing capacity) raised trimming probabilities. In contrast, increasing both AT and GC nucleotides 3' of the motif (which increases absolute position) reduced trimming probabilities for both gene types (Figure 4.3A).

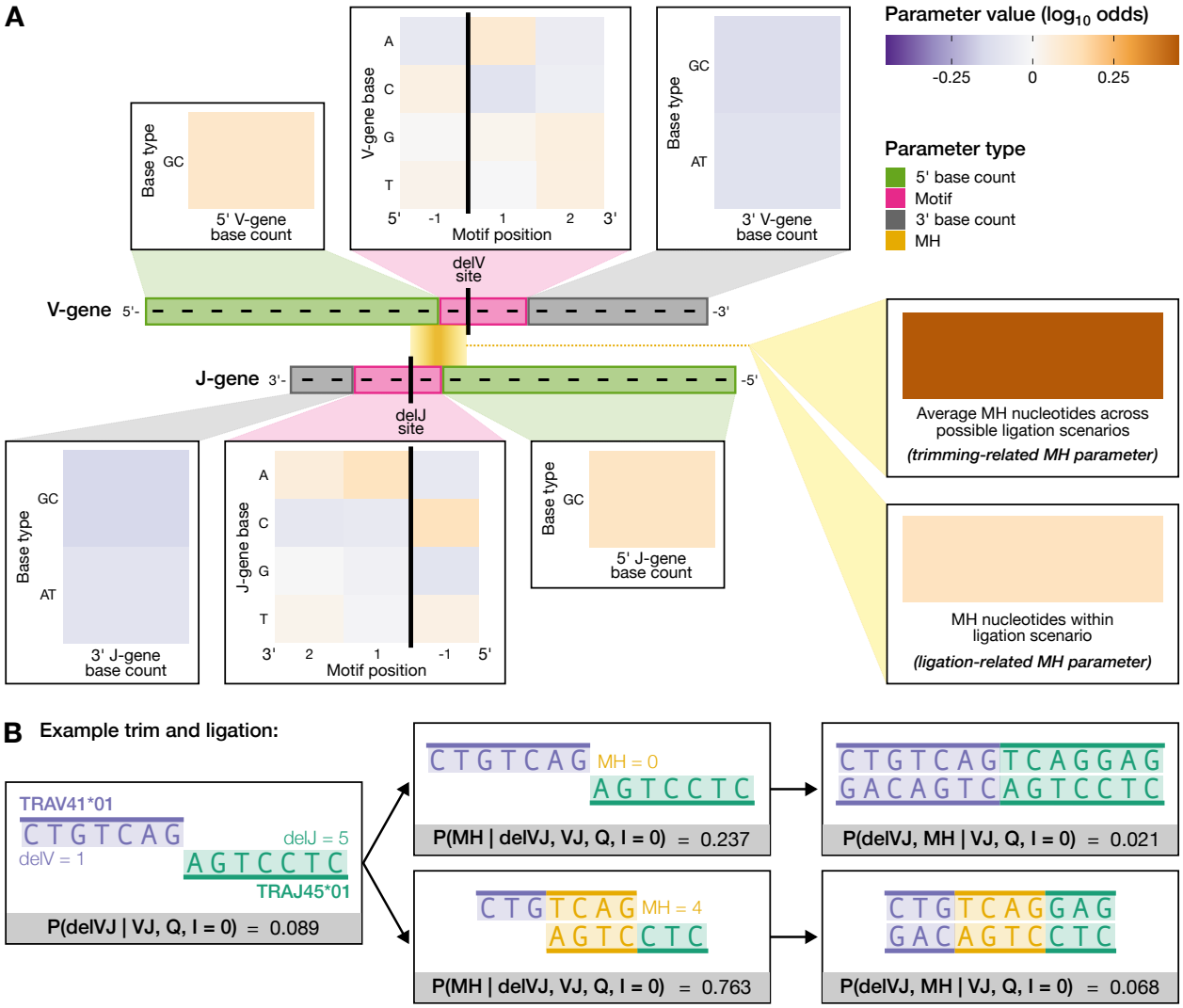
A
Parameter value (log10 odds)
-0.25
0
0.25
Parameter type
5' base count
Motif
3' base count
MH
Base type
GC
5' V-gene base count
V-gene base
A
C
G
T
5'
-1
1
2
3'
Motif position
Base type
GC
AT
3' V-gene base count
delV site
V-gene 5'-
-3'
J-gene 3'-
-5'
delJ site
Average MH nucleotides across possible ligation scenarios
(trimming-related MH parameter)
MH nucleotides within ligation scenario
(ligation-related MH parameter)
Base type
GC
AT
3' J-gene base count
J-gene base
A
C
G
T
3'
2
1
-1
5'
Motif position
Base type
GC
5' J-gene base count
B
Example trim and ligation:
TRAV41*01
CTGTCAG
delV = 1
delJ = 5
AGTCCTC
TRAJ45*01
P(delVJ | VJ, Q, I = 0) = 0.089
CTGTCAG MH = 0
AGTCCTC
P(MH | delVJ, VJ, Q, I = 0) = 0.237
CTGTCAGTCAGGAG
GACAGTCAGTCCTC
P(delVJ, MH | VJ, Q, I = 0) = 0.021
CTGTCAG MH = 4
AGTCCTC
P(MH | delVJ, VJ, Q, I = 0) = 0.763
CTGTCAGGAG
GACAGTCCTC
P(delVJ, MH | VJ, Q, I = 0) = 0.068
C
relative importance
0.0
0.1
0.2
3' V-gene GC count
3' J-gene GC count
3' J-gene AT count
trimming-related MH
3' V-gene AT count
5' J-gene GC count
5' V-gene GC count
5' J-motif C
ligation-related MH
3'-pos1 J-motif A
5' J-motif G
3'-pos1 J-motif G
5' V-motif A
3'-pos1 V-motif C
5' J-motif A
3'-pos1 V-motif A
3'-pos2 J-motif C
3'-pos1 J-motif C
3'-pos2 J-motif A
5' V-motif C
5' J-motif T
3'-pos2 V-motif A
3'-pos2 V-motif G
3'-pos2 V-motif T
5' V-motif T
3'-pos1 J-motif G
3'-pos2 J-motif T
3'-pos1 V-motif G
3'-pos1 J-motif T
3'-pos2 J-motif G
5' V-motif G
3'-pos1 V-motif T

Figure 4.3: Although sequence-based parameters such as gene-specific trimming motifs and base counts contribute meaningfully to predicting trimming and ligation probabilities, the extent of germline-encoded microhomology (MH) between sequences exerts a strong effect, especially in increasing trimming probabilities. **(A)** Illustration of sequence features and their alignment with an arbitrary V- and J-gene pair at example trimming sites, which correspond to the number of nucleotides deleted from each gene (represented by the random variables delV and delJ), along with inferred model parameters. V- and J-gene trimming motif parameters (pink) reflect the influence of adjacent nucleotides on trimming probabilities. Trimming motif positions are indexed relative to the inferred trimming site for each gene, with negative indices indicating positions 5' of the trimming site and positive indices indicating positions 3'. V- and J-gene base count parameters (green and gray) reflect the influence of upstream and downstream AT and GC nucleotide composition on trimming probabilities. Specifically, we find that an increase in GC nucleotides 5' of the motif increases trimming probabilities, while an increase in AT or GC nucleotides 3' of the motif decreases them. The model excludes 5' AT nucleotide counts. MH between sequences (gold box) strongly influences both trimming and ligation probabilities, with a larger positive effect on trimming. Black vertical lines indicate example trimming sites. Each parameter represents the change in $\log_{10}$ odds of trimming or ligating due to an increase in the feature value, assuming all other features are held constant. **(B)** Our model demonstrates that increasing MH generally raises trimming and ligation probabilities, as shown in example scenarios for the most frequently used gene pair, TRAV41*01 (purple) and TRAJ45*01 (green). In the bottom row, four nucleotides of MH (gold) result in a most probable trimming and ligation scenario (left and middle boxes) with a joint probability of 0.068 (right box). In contrast, the top row shows the same trimmed sequences ligating with zero MH, leading to a lower joint probability of 0.021. Trimming and ligation probabilities are inferred across trimming scenarios (delVJ) and ligation scenarios (MH) for a V-J gene pair (VJ), yielding sequence productivity (Q) with zero N-insertions $(\mathrm{I} = 0)$. **(C)** Parameters for 3' AT and GC base counts have the highest relative importance (gray), followed by the trimming-related microhomology parameter (yellow). Relative importance was calculated using a model trained with standardized features, where the absolute values of parameter estimates indicate their contribution to the model.

Finally, we examined the relative effect sizes and importances of these sequence-level parameters to identify the most influential factors affecting trimming and ligation outcomes. The strongest positive effects were observed for trimming-related germline-encoded microhomology effects $(\text{parameter} = 0.4484)$, followed by the presence of a C nucleotide immediately 5' of the J-gene trimming site $(\text{parameter}_J = 0.1308)$, and ligation-related germline-encoded

microhomology effects (parameter = 0.1272) (Figure 4.3A and Figure E.9). In contrast, the most negative effects were an increase in GC nucleotides 3' of the motif for both V- and J-genes ($\text{parameter}_V = -0.1321$, $\text{parameter}_J = -0.1512$), the presence of a C nucleotide 3' of the V-gene trimming site ($\text{parameter}_V = -0.1049$), and an increase in AT nucleotides 5' of the motif for both V- and J-genes ($\text{parameter}_V = -0.1095$, $\text{parameter}_J = -0.1030$). P-values corresponding to each of these effects were smaller than machine tolerance ($p \simeq 0$).

To evaluate the relative importance of model parameters, we trained our model using standardized features which ensure that parameter estimates directly reflect their relative importance to the model. This analysis revealed that the counts of AT and GC nucleotides 3' of the motif for both V- and J-genes were the most influential, closely followed by the parameter representing trimming-related microhomology effects (Figure 4.3C). Parameters corresponding to the counts of GC nucleotides 5' of the motif and ligation-related microhomology effects were also identified as relatively important.

4.3.2 Germline-encoded microhomology significantly improves model fit for predicting trimming and ligation across other receptor loci and sequence types

To further assess the importance of incorporating germline-encoded microhomology-related parameters for accurately predicting trimming and ligation probabilities, we compared the performance of a full model, which includes germline-encoded microhomology, motif, and 5' and 3' base count terms, to models lacking specific terms. All models were trained using the non-productive TCRα training dataset and the parameters were held constant for subsequent analyses.

We began by evaluating model performance on the training dataset. The full model showed a substantially lower expected per-sequence log loss compared to the model without germline-microhomology-related parameters, indicating a better fit to the data (Figure 4.4A). This improvement was validated by a likelihood ratio test (LRT), which confirmed the sta-

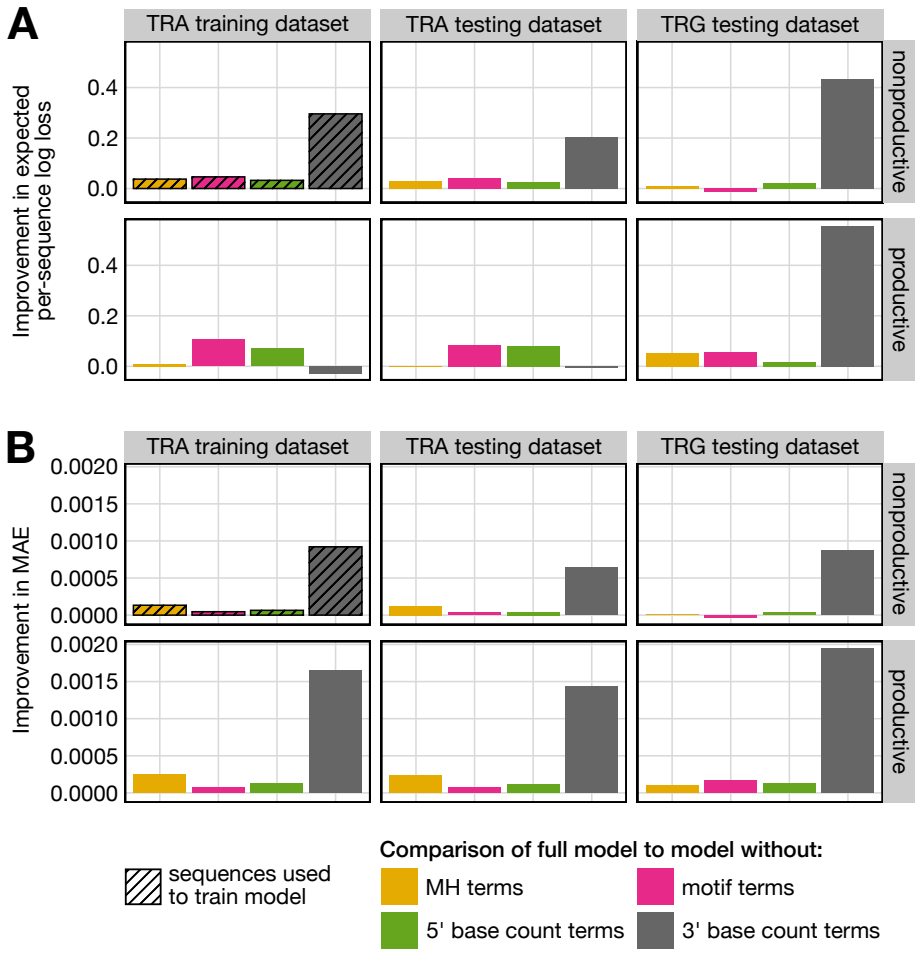

Figure 4.4: **(A)** Improvement in expected per-sequence log loss for the full model, which includes germline-encoded microhomology (MH) terms, motif terms, and 5' and 3' base count terms, compared to models without specific terms across both productive and nonproductive sequences from multiple datasets. Improvement is the negative difference in log loss, with negative values indicating a relatively worse fit and positive values indicating a relatively better fit for the full model. Including MH terms improves log loss across all datasets, except for productive sequences from the TCRα testing dataset, where no change in loss was observed. **(B)** Improvement in mean absolute error (MAE) across the same models and datasets. Improvement is the negative difference in MAE, with negative values indicating relatively lower predictive accuracy and positive values indicating relatively higher predictive accuracy for the full model. Including MH terms consistently improves MAE across all datasets. All models were trained using nonproductive sequences from the TCRα training dataset (hatched boxes), with parameters held constant ("frozen") before calculating log loss and MAE across datasets.

tistical significance of including germline-encoded microhomology terms (LRT statistic = 93754.84; p-value less than machine tolerance, $p \simeq 0$). The full model also exhibited higher

predictive accuracy, as indicated by a lower mean absolute error (MAE = 0.00468) compared to the model without germline-encoded microhomology terms (MAE = 0.00481). We repeated this analysis across models lacking other parameter types and found that the full model consistently outperformed them, exhibiting lower expected per-sequence log loss and MAE in each case (Figure 4.4). Recall that the 5' base count parameters capture potential sequence-breathing effects by reflecting preferences for GC content upstream of the motif, while the 3' base count parameters capture both preferences for the absolute position of the trimming site and sequence-breathing effects related to AT and GC content downstream of the motif. Among the individual parameter types, the 3' base count terms had the largest impact, leading to the greatest improvement in both log loss and MAE. Microhomology and motif terms contributed the second-largest improvements in MAE and log loss, respectively. That is, the absolute position of the trimming site, represented by the 3' base count terms, had the strongest influence, while the local nucleotide context at the trimming site (captured by motif terms) and the extent of germline-encoded microhomology between the trimmed and ligated sequences also provided positive contributions, though to a lesser extent. Sequence-breathing capacity upstream of the trimming site, reflected by the 5' base count terms, improved log loss and MAE as well, but had a smaller overall effect compared to the other parameters.

Using frozen coefficients from our models trained on nonproductive TCRα sequences without N-insertions, we can also infer trimming and ligation probabilities for productive sequences or sequences from other receptor loci. However, because our models are specifically designed for sequences lacking N-insertions——since N-insertions complicate ligation pattern analysis due to their unknown nucleotide composition prior to ligation——its inferences are limited to such sequences. To evaluate model performance, we tested all models on both productive and nonproductive sequences from independent TCRα and TCRγ datasets. The full model consistently demonstrated superior predictive accuracy, achieving lower expected

per-sequence log loss and mean absolute error (MAE) compared to alternative models (Figure 4.4).

In most datasets, the inclusion of 3' base count terms continued to have the strongest impact on improving model fit and predictive accuracy. However, there were two notable exceptions in log loss calculations for productive sequences from the TCRα training and testing datasets. In these cases, including 3' base count terms, which capture effects related to the absolute positioning of the trimming site, negatively affected log loss. Since productive sequences are subject to selection-related effects that may alter preferences for trimming site positioning, the 3' base count terms learned from nonproductive sequences——where these selection effects are absent——may be less effective for predicting trimming in productive sequences. Nevertheless, the inclusion of 3' base count terms still improved MAE in these cases, despite the negative impact on log loss. This discrepancy may stem from log loss being more sensitive to outliers than MAE.

The inclusion of microhomology terms also improved model fit and predictive accuracy across most datasets, consistently providing the second-largest improvement in MAE. Notably, even when applied to productive sequences from the TCRα testing set——despite these sequences not being included in training and having skewed recombination statistics due to selection——the full model outperformed the model without microhomology terms in MAE, although the log loss values were similar. This suggests that while the inclusion of microhomology terms improves log loss across datasets, their most pronounced impact is on MAE. Overall, these consistent findings across other receptor loci (i.e. TCRγ) and sequence types (i.e. productive sequences) highlight the biological significance of germline-encoded microhomology in accurately modeling trimming and ligation scenarios.

4.3.3 Accounting for germline-encoded microhomology affects sequence annotation

Given the significant role of germline-encoded microhomology in predicting trimming and ligation scenarios across TCRα and TCRγ receptor loci, we wanted to evaluate how germline-encoded microhomology parameterization influences sequence annotation. Recall that sequence annotation involves assigning a specific V(D)J recombination annotation, which describes the associated trimming, insertion, and ligation scenarios, to an observed sequence. In earlier sections, we examined the joint probabilities of trimming and ligation scenarios *for V-J gene pairs*, which represent the normalized probability of each trimming and ligation scenario within the complete set of possibilities for a given gene pair. Here, we shift our focus to V(D)J recombination annotation probabilities *for individual observed sequences*, which represent the normalized probability of each V(D)J recombination annotation within all possible annotations for a given sequence. Since we are analyzing sequences without N-insertions, each V(D)J recombination annotation corresponds directly to a trimming and ligation scenario, allowing us to use our inferred joint trimming and ligation scenario distributions to calculate the corresponding V(D)J recombination annotation probabilities. In this analysis, we compare the V(D)J recombination annotation probabilities and rankings between two models: (1) the full model, which includes germline-encoded microhomology, motif, and 5' and 3' base count terms, and (2) a version of the model that excludes germline-encoded microhomology terms.

As expected from our earlier results, accounting for germline-encoded microhomology effects substantially alters annotation probabilities and their rankings for sequences with multiple possible annotations. In total, 9.2% of all sequences lacking N-insertions have a different top-ranked annotation when using the model that parameterizes germline-encoded microhomology compared to the model that does not. These sequences represent the subset where microhomology-related effects could be inferred given our model setup and were actually detected.

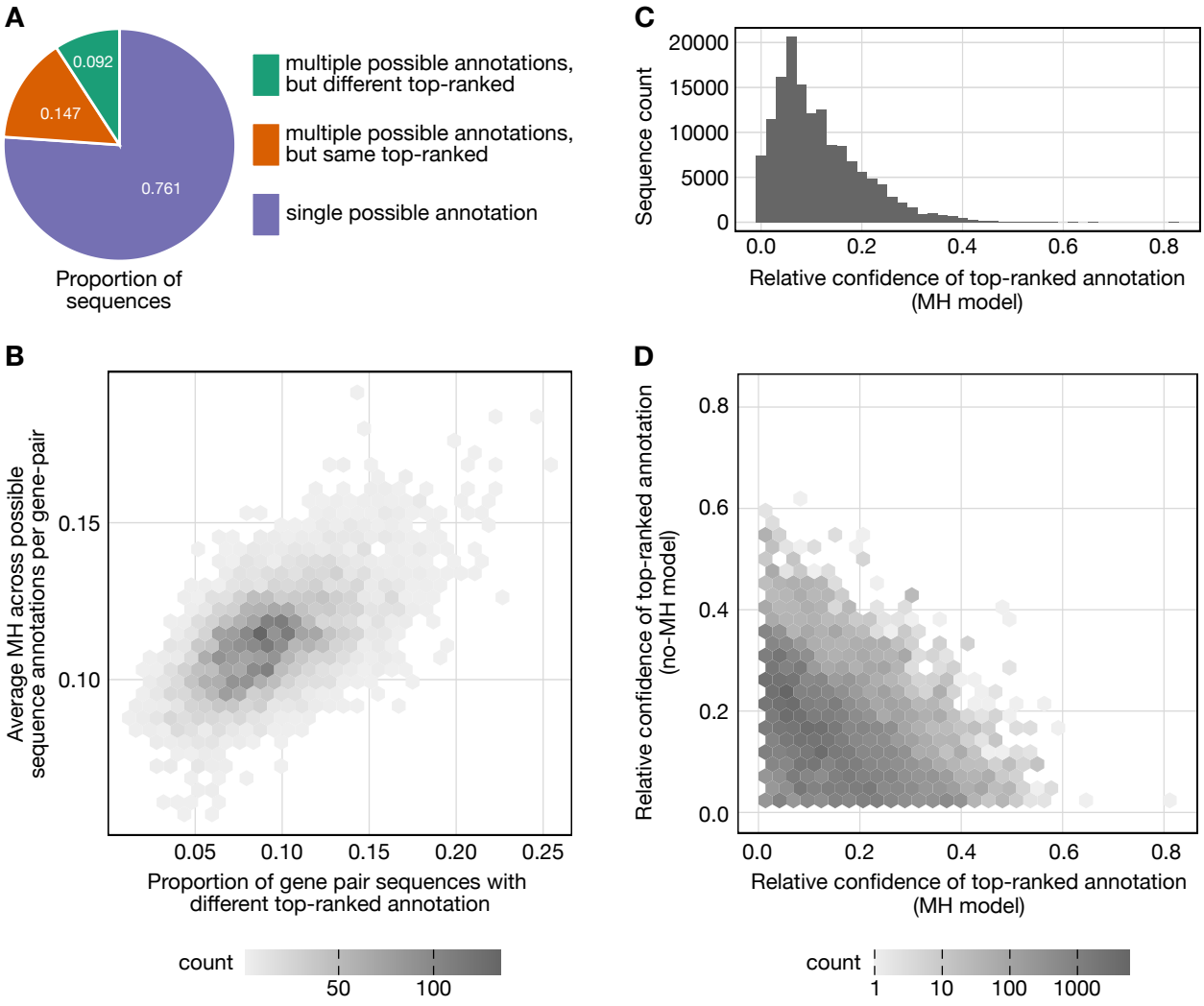
A
0.092
0.147
0.761
multiple possible annotations, but different top-ranked
multiple possible annotations, but same top-ranked
single possible annotation
Proportion of sequences
C
Sequence count
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
Relative confidence of top-ranked annotation (MH model)
B
Average MH across possible sequence annotations per gene-pair
0.15
0.10
0.05
0.10
0.15
0.20
0.25
Proportion of gene pair sequences with different top-ranked annotation
count
50
100
D
Relative confidence of top-ranked annotation (no-MH model)
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
Relative confidence of top-ranked annotation (MH model)
count
1
10
100
1000

Figure 4.5: Accounting for germline-encoded microhomology (MH) in models used for sequence annotation substantially alters both V(D)J recombination annotation probabilities and rankings. For each sequence, the top-ranked annotation was determined separately using the model that parameterizes germline-encoded microhomology (MH model) and the model that does not (no-MH model). **(A)** Proportions of sequences categorized by whether they have single or multiple possible V(D)J recombination annotation scenarios and whether their top-ranked V(D)J recombination annotation differs between the MH and no-MH models. The majority of sequences lacking N-insertions have only one possible annotation, meaning microhomology-related effects could not influence their ranking. These statistics are specific to sequences without N-insertions, where the ability to detect microhomology-related annotation effects is more limited due to the generally lower number of possible annotations in this subset. If these effects were quantified in sequences containing N-insertions, a larger fraction would likely exhibit differences in top-ranked annotations when accounting for microhomology. **(B)** Correlation between the proportion of sequences with differing top-ranked annotations (between the two models) and the average germline-encoded microhomology across potential annotations per sequence for each V-J gene pair. This highlights how germline-encoded microhomology may influence ranking changes across V-J gene pairs. **(C)** Distribution of relative confidence for the top-ranked annotation using the MH model. Relative confidence is defined as the absolute difference in annotation probabilities using the MH model, comparing the top-ranked annotation from the MH model to the top-ranked annotation from the no-MH model for each sequence. **(D)** Comparison of relative annotation confidence for each sequence between the two models. Relative confidence for a model is calculated as the absolute difference in annotation probabilities (from that model) for the two top-ranked annotations identified by the MH and no-MH models. Most sequences exhibit substantial shifts in relative confidence between the two models, highlighting large model-driven changes in sequence annotations, even when one or both models show high relative confidence.

However, our model can only detect germline-microhomology-related effects in sequences that allow for such inference. The majority of N-insertion-lacking sequences (76.1%) have only one possible annotation, meaning microhomology-related effects could not influence their ranking. Among sequences with multiple possible annotations—where microhomology-related effects could, in principle, be detected—38.3% exhibit a change in the top-ranked annotation.

Since sequences without N-insertions tend to have fewer possible annotations than those

with N-insertions, the ability to detect microhomology-related annotation effects is more limited in this subset. If we were to quantify these effects in sequences containing N-insertions, we might expect a larger fraction of sequences to have a different top-ranked annotation when microhomology is accounted for.

For sequences lacking N-insertions where microhomology does influence annotation rankings, the magnitude of this effect appears to depend on the amount of germline-encoded microhomology present. Specifically, as the average germline-encoded microhomology across potential annotations increases for a given sequence within a V-J gene pair, the proportion of gene pair sequences with differing top-ranked annotations between the two models also increases (Figure 4.5B). We quantified the significance of this relationship using Pearson's correlation, which revealed a moderately positive correlation ($r = 0.5877$; p-value smaller than machine tolerance, $p \simeq 0$). For some V-J gene pairs, parameterizing germline-encoded microhomology has a particularly pronounced impact on sequence annotation. For example, sequences involving TRAV38-1*04 and TRAJ22*01 show a striking difference in annotation predictions between models, with 25.13% of sequences exhibiting different top-ranked annotations. This effect may be driven by the relatively high germline-encoded microhomology content across annotations for these sequences, averaging 0.1697 nucleotides compared to the overall average of 0.1144 nucleotides across all V-J gene pairs.

Given that parameterizing germline-encoded microhomology leads to different top-ranked annotations for many sequences, we next quantified the relative confidence of these rankings. To explore this, we compared the annotation probabilities assigned by the model with germline-encoded microhomology terms for the top-ranked annotations from the microhomology model and the no-microhomology model for each sequence. We define the relative confidence of a top-ranked annotation as the absolute difference in annotation probabilities compared to the top-ranked annotation from the other model. On average, for sequences containing a different top-ranked annotation between the two models, we find that the rel-

ative confidence of the top-ranked annotation for the microhomology model is 0.1140 (Figure 4.5C). This means that, on average, the top-ranked annotation from the microhomology model is 11.4% more probable than the top-ranked annotation from the no-microhomology model, based on probabilities assigned by the microhomology model.

Additionally, we examined the relative confidence levels of the top-ranked annotations from both models. If germline-encoded microhomology merely resolved ties between competing annotations, we would expect minimal relative confidence in the top-ranked annotation using the model lacking microhomology terms, with larger relative confidence observed for the model containing microhomology terms. However, our findings indicate substantial shifts in relative confidence across models for most sequences (Figure 4.5D), with data points widely distributed rather than clustering near the axes. For instance, even when the model lacking microhomology terms has high confidence in its top-ranked annotation relative to the top-ranked annotation derived from the model containing microhomology, a similar flip in confidence is often observed when switching models. This effect suggests that parameterizing germline-encoded microhomology leads to meaningful changes in the annotation ranking landscape, potentially altering the biological interpretation of many sequences.

Beyond sequence annotation probabilities and rankings, we were interested in exploring whether germline-microhomology-related effects had practical implications for sequence generation probabilities, which are often used to characterize immune repertoire sequences. Our analysis revealed a small but consistent difference in sequence generation probabilities between the model that includes microhomology effects and the one that does not (Figure E.10). Notably, these differences correlate with the average number of microhomologous nucleotides across all possible annotation scenarios for a given sequence (Figure E.11), likely reflecting the influence of microhomology on sequence generation. These results suggest that incorporating microhomology effects into generative models of immune repertoire sequencing could enhance their biological relevance and improve their utility as negative controls, a

common application in the literature [20, 123, 124].

4.4 Discussion

Previous *in vitro* experiments have suggested that germline-encoded microhomology plays a significant role in biasing key V(D)J recombination processing steps, such as trimming and ligation. However, these findings do not fully elucidate the importance of germline-encoded microhomology in shaping *in vivo* recombination outcomes in humans. In this chapter, we use statistical inference on previously-published high-throughput human TCR repertoire data [118, 119] to assess whether germline-encoded microhomology influences V(D)J recombination in humans with intact recombination machinery. Our probabilistic modeling framework quantifies how sequence-level features, particularly germline-encoded microhomology, impact trimming and ligation decisions during V(D)J recombination. We find that (1) germline-encoded microhomology significantly increases trimming and ligation event probabilities such that each additional nucleotide of microhomology increases the odds of a trimming event by 181% and the odds of a ligation event by 34%, (2) incorporating germline-encoded microhomology terms significantly enhances model fit for predicting trimming and ligation across multiple receptor loci and sequence types, and (3) accounting for microhomology when inferring V(D)J recombination annotations alters annotation probabilities and rankings, leading to a qualitatively different top-ranked annotation for 38.2% of sequences with multiple possible annotations.

Our results reveal that germline-encoded microhomologous nucleotides between gene ends significantly increase the probabilities of ligation events, aligning with previous *in vitro* evidence suggesting that germline-encoded microhomology guides ligation [26, 44–46, 109–111]. While much of the previous experimental focus has been on terminal microhomology (present at gene ends), many gene pairs lack terminal microhomology but have in-

terior regions of microhomology. It has been proposed that trimming can expose these interior regions, which then guide ligation through germline-microhomology-mediated processes [46, 110]. Our findings support this, as germline-encoded microhomology appears to have a stronger effect on trimming than on ligation, likely due to the dependence of ligation options on prior trimming choices. Because this analysis focuses on sequences without N-insertions——allowing us to directly identify germline-microhomology-mediated ligation events——the observed strength of these effects may be amplified compared to analyses that include all sequences. Notably, when analyzing sequences with N-insertions——where ligation is not mediated by germline-encoded microhomology——we still observe that germline-encoded microhomology influences trimming, though less strongly, suggesting a more complex role for germline-encoded microhomology in V(D)J recombination beyond its involvement in ligation.

In addition to germline-microhomology-related parameters, our modeling framework included sequence-level parameters designed to capture the effects of local nucleotide context, absolute trimming site positioning, and sequence breathing capacity. These parameters, except for the germline-microhomology-related ones, were introduced in our previous analyses of trimming patterns for individual V- and J-gene sequences [3]. Our current results were consistent with those earlier findings. Specifically, parameters capturing the effects of absolute trimming site positioning and sequence breathing capacity downstream of the trimming site had the most substantial impact on improving overall model fit and predictive accuracy, showing the largest negative effect sizes on trimming and ligation choices. This pattern held when evaluating model performance and accuracy with sequences from different receptor loci and productivity types (e.g. TCRγ sequences and productive TCRα sequences), highlighting the influence of germline-encoded microhomology on recombination decisions. An important next step could involve investigating microhomology-related effects in other receptor loci that are more challenging to study, such as TCRβ, which has less sequence diversity be-

tween joining genes, and *IGH*, which undergoes post-recombination somatic hypermutation.

Beyond its intrinsic interest, germline-encoded microhomology has significant practical implications. In addition to influencing trimming and ligation probabilities, we found that parameterizing germline-microhomology-related effects leads to shifts in V(D)J recombination annotation probabilities and rankings, as well as sequence generation probabilities. These shifts often corresponded to large changes in the relative confidence of annotation rankings when comparing models that incorporate germline-encoded microhomology with those that do not. Such changes could meaningfully alter the annotation and sequence generation probability landscape, potentially impacting the biological interpretation of many sequences.

Our analysis was restricted to sequences lacking N-insertions, which tend to have fewer possible annotations per sequence compared to those with N-insertions. Because microhomology-related annotation effects can only be detected in sequences with multiple possible annotations, the ability to observe these effects is inherently more restricted in this subset. If we were to quantify these effects in sequences containing N-insertions, a larger fraction would likely exhibit differences in top-ranked annotations when microhomology is accounted for. Despite these findings, to our knowledge, all widely used V(D)J recombination annotation software [9–11] and generative models of immune repertoire sequencing data [11] do not account for germline-encoded microhomology or consider annotations that incorporate germline-encoded microhomologous nucleotides.

Our work has several limitations. First, we rely on non-productive rearrangements as a proxy for pre-selection recombination statistics, as is common in the literature [11, 15, 19, 20]. Non-productive sequences are sequenced as part of the repertoire when they coexist within a cell expressing a productive rearrangement that has passed the selection process. While we are not aware of any mechanism that could correlate nonproductive and productive rearrangements within a single cell, nor of any evolutionary pressures acting to minimize the

frequency of nonproductive rearrangements, we acknowledge that the repertoire of nonproductive rearrangements may not perfectly reflect the pre-selection repertoire. Nevertheless, we assume that recombination events are independent and that nonproductive rearrangements reasonably approximate the recombination statistics of the repertoire before selection. Second, our analysis excluded sequences with N-insertions, allowing us to use known germline V- and J-gene sequences to identify regions of germline-encoded microhomology and potential germline-microhomology-mediated ligation events. N-insertions complicate the analysis because the identities of inserted nucleotides are unknown, and their presence suggests germline-encoded microhomology did not guide ligation. Because the presence or absence of N-insertions may affect the probability of successful ligation, excluding these sequences could shift the distribution of observed trimming and ligation events. Consequently, the germline microhomology effects that we have inferred may not extend to sequences with N-insertions. Future work could explore insertion-dependent microhomology dynamics in sequences containing N-insertions, but doing so would require assumptions about and integration over latent variables such as N-insertion identities prior to ligation, making this analysis challenging if using repertoire sequencing data. Relatedly, future work could also investigate how microhomology influences gene usage inference during V(D)J recombination annotation.

Despite the clear role of germline-encoded microhomology in biasing V(D)J recombination events and influencing V(D)J recombination annotation inference, no probabilistic models incorporating microhomology have been developed, to our knowledge. Future work could integrate microhomology-related dependencies into existing probabilistic frameworks like IGoR [11], which currently models dependencies between recombination events such as V- and J-gene choice, V-gene choice and V-gene deletions, and J-gene and J-gene deletions for TCRα sequences. To explicitly account for microhomology, additional dependencies would need to be introduced between V- and J-gene deletions, V-gene choice and J-gene deletion,

and J-gene choice and V-gene deletion, along with incorporating new parameters to capture the sharing of germline-encoded microhomologous nucleotides. However, this approach could be challenging due to the large number of parameters required and the corresponding need for large datasets to adequately train the model. Alternatively, one could replace junctional processing event terms (such as those related to trimming and insertion) within IGoR with a more generalized model of junctional processing that incorporates microhomology, such as the model presented here. This modification would substantially reduce the number of required parameters, potentially balancing model complexity with practicality, although it might limit the ability to capture gene-specific processing profiles. Other more advanced approaches, such as combining simulation with deep learning, could also be explored to account for microhomology in V(D)J recombination annotation inference.

In summary, our findings demonstrate that germline-encoded microhomology plays a significant role in trimming and ligation choices during V(D)J recombination, underscoring the importance of accounting for germline-encoded microhomology effects when predicting recombination outcomes and annotating sequences. By advancing our understanding of the influence of germline-encoded microhomology in human V(D)J recombination, these results provide another step toward uncovering how this process generates diverse receptors that support a robust immune response in humans.

Chapter 5

CONCLUSIONS

Understanding the intricate processing steps underlying V(D)J recombination is essential for elucidating how immune repertoire diversity is generated and maintained to support robust immune responses to disease and vaccination. This dissertation advances our understanding of the mechanisms and variability of V(D)J recombination by examining genetic influences, nucleotide trimming dynamics, and the role of sequence microhomology in shaping recombination outcomes. Below, I summarize the key findings and propose future research directions.

5.1 To what extent does V(D)J recombination vary across individuals?

This dissertation demonstrates that V(D)J recombination probabilities vary significantly across individuals, influenced by genetic variation in recombination-related genes. In Chapter 2, I identified genetic variants associated with biases in nucleotide trimming and N-insertions. While these variants have modest effects (e.g., approximately 0.1 nucleotide differences), they underscore how genetic variation can shape junctional diversification. Observed ancestry-associated differences in N-insertion patterns further suggest that inherited genetic variation influences recombination probabilities. Additionally, I confirmed and extended prior findings on genetic determinants of TCR gene usage, observing substantial effects (e.g., up to a 4% difference in repertoire-level gene usage). Together, these findings

illuminate how genetic variation may impact TCR repertoire generation and diversity.

Despite these associations, the direct mechanistic effects of the identified variants remain unclear. Experimental validations, such as gene expression analyses or functional recombination assays, are necessary to establish causality. Mechanistic modeling, as applied in Chapter 3, also offers a promising avenue for exploring how these variants influence nucleotide-level recombination processes. The increasing availability of large immune receptor repertoire datasets provides an opportunity to continue applying such models at high resolution.

Understanding the functional consequences of V(D)J recombination biases is essential for elucidating their contributions to disease susceptibility. Future research could explore correlations between TCR repertoire composition and immune exposures or disease states (e.g., as in [50]) while incorporating the genetic biases identified here to refine our understanding of immune variability. These investigations would benefit from integrating T and B cell data. For B cells, the influence of genetic variation on V(D)J recombination probabilities remains largely unexplored. Paired B cell receptor repertoire sequencing and whole-genome sequencing/genotyping, with careful control for confounding factors such as immune exposure history and somatic hypermutation, will be crucial for addressing this gap.

5.2 How are nucleotides trimmed during V(D)J recombination in humans?

Molecular experiments have provided critical insights into nucleotide trimming during V(D)J recombination. However, direct studies in humans remain limited, particularly those examining the mechanisms by which the Artemis nuclease may trim nucleotides. In Chapter 3, I address this gap using model-based statistical inference on high-throughput repertoire sequencing data to investigate how sequence-level features influence nucleotide trimming probabilities in humans.

This approach identified universal sequence-level features that predict trimming probabilities across all adaptive immune receptor loci with high accuracy. Notably, trimming was more likely to occur at DNA positions closer to the end of the sequence, likely reflecting increased accessibility. I also found evidence that double-stranded DNA may need to "breathe" to enable trimming. Additionally, I identified a sequence motif preferentially targeted for trimming, independent of breathing and position, suggesting a role for sequence-specific interactions with nucleases.

Importantly, I demonstrated that genetic variation in Artemis, identified in Chapter 2, modulates these inferred parameters, offering new mechanistic insights into how genetic differences influence trimming probabilities. These findings refine our understanding of the nucleotide trimming process in humans and establish a foundation for investigating individual differences in trimming mechanisms.

Future experimental studies are essential to validate these findings. For example, in vitro experiments could test whether specific DNA motifs influence Artemis-mediated trimming or whether end-proximity and breathing effects correspond to measurable changes in DNA accessibility or stability. Establishing a feedback loop between statistical modeling and experimental validation will be critical for furthering our understanding of the molecular mechanisms underlying the trimming process.

5.3 Does sequence microhomology bias V(D)J recombination outcomes?

Previous studies have suggested that microhomology biases V(D)J recombination, but its role in shaping recombination outcomes in humans has remained unclear. In Chapter 4, I developed a probabilistic model of trimming and ligation that incorporates microhomology to investigate these patterns. Using high-throughput human repertoire sequencing data, this model revealed that microhomology significantly biases recombination outcomes across adap-

tive immune receptor loci by influencing both trimming and ligation steps. Additionally, I found that accounting for microhomology in V(D)J recombination annotation also substantially alters annotation probabilities and rankings, affecting the biological interpretation of many immune receptor sequences in downstream analyses. These findings highlight both the mechanistic importance of microhomology and its practical significance for improving annotation accuracy.

While this study focused on trimming and ligation, microhomology likely influences other V(D)J recombination processes, such as N-insertion. Because standard repertoire-based sequencing captures receptor sequences only after V(D)J recombination is complete, the identities of nucleotides inserted prior to ligation are not observed. As a result, quantifying the influence of microhomology on N-insertion using these data may be more challenging and require a simulation-based probabilistic modeling approach to address this added complexity.

Building on these results, a comprehensive model of V(D)J recombination that incorporates microhomology across all processing steps should be developed to improve the accuracy of V(D)J recombination annotation inference. Existing frameworks like IGoR [11] could be adapted by replacing current junctional processing parameters (e.g., gene-specific trimming and insertion profiles) with a generalized model integrating microhomology effects, similar to the model introduced in this dissertation. Advanced methods combining simulation with deep learning could also be explored to account for microhomology in annotation inference.

5.4 Final remarks and implications

This dissertation provides key insights into the molecular mechanisms of V(D)J recombination by exploring individual variability, the nucleotide trimming process, and the role of microhomology. These findings lay a foundation for future explorations into how recombina-

tion processes shape immune diversity and impact immune responses. A robust mechanistic understanding of V(D)J recombination variability will be critical for the future development of personalized immune-based therapies and vaccine design. This work represents a significant step toward unraveling the complex processes underlying immune receptor generation and their implications for human health.

Appendix A

SUPPLEMENTARY TABLES AND FIGURES FOR CHAPTER 2

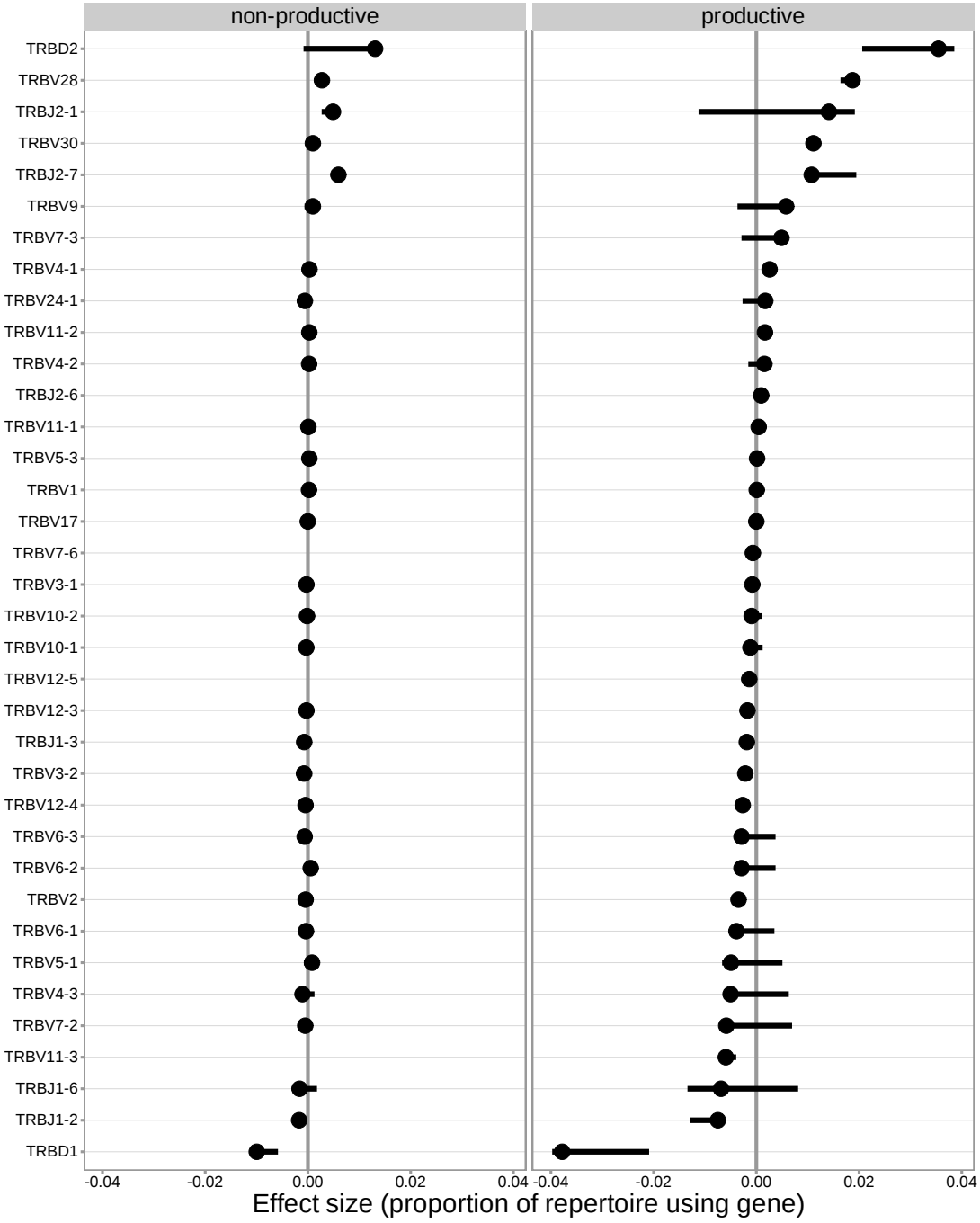

Figure A.1: For the significantly associated *TRB* locus SNPs, the median association effect magnitude was largest for the expression of TRBD1 followed by the expression of TRBD2 and the expression of TRBV28 all in productive TCRs. The median association effect magnitude for each gene is shown by each point and the interquartile range of the association effect sizes for each gene is given by each black horizontal line.

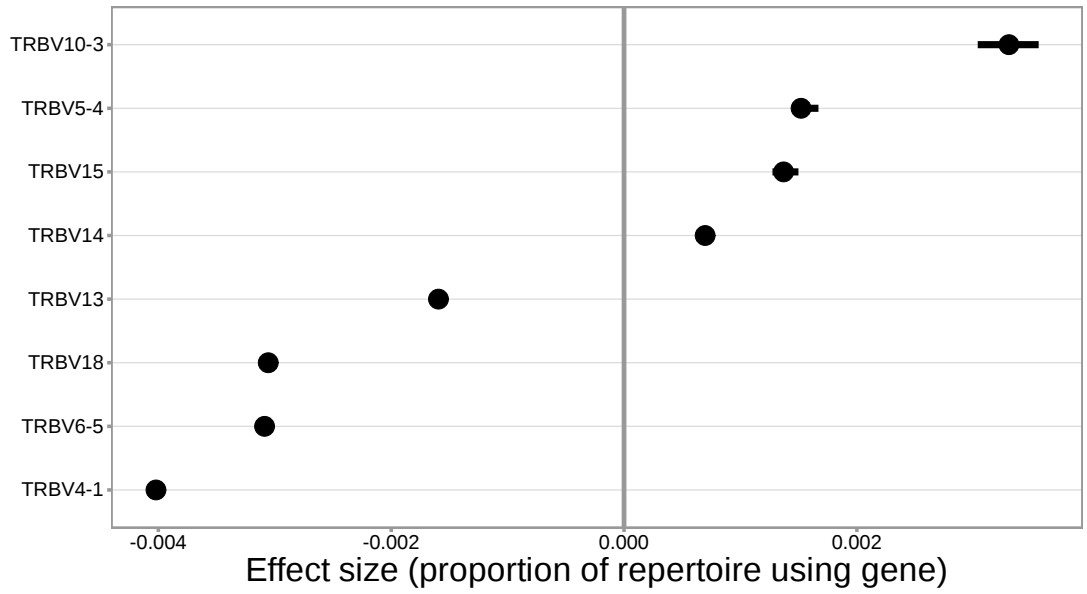

Figure A.2: For the significantly associated MHC locus SNPs, the median association effect magnitude was largest for the expression of TRBV4-1 followed by the expression of TRBV10-3. The median association effect magnitude for each gene is shown by each point and the interquartile range of the association effect sizes for each gene is given by each black horizontal line.

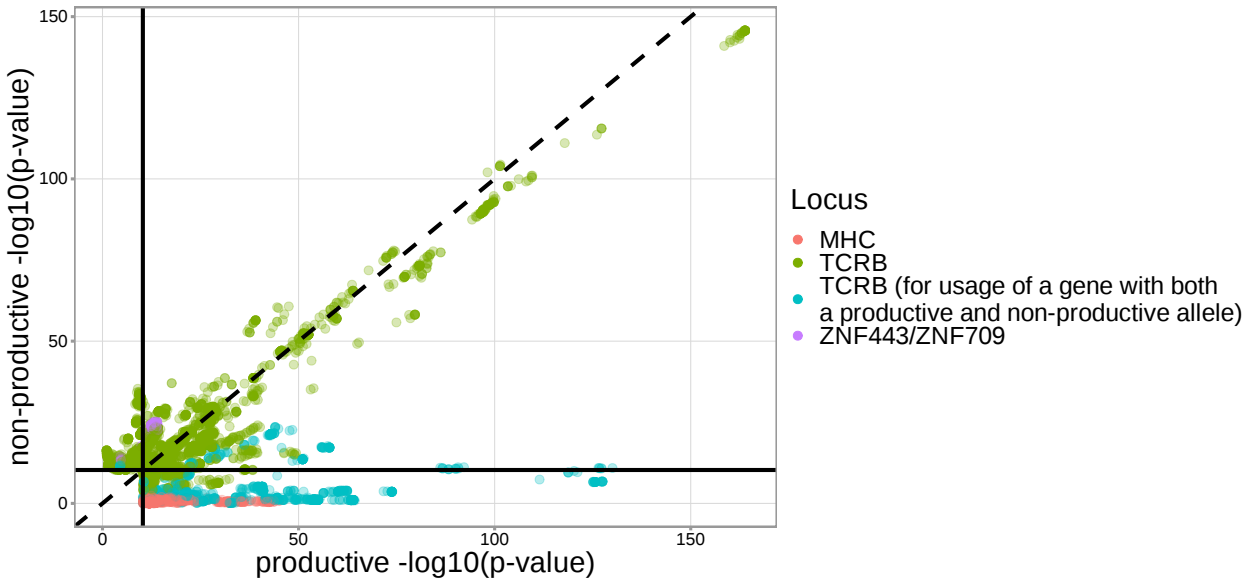

Figure A.3: The majority of significantly associated *TRB* locus SNPs had similar gene usage association P-values between non-productive and productive TCRs, but significantly associated MHC locus SNPs were only significant for gene usage of productive TCRs. Notably, the majority of *TRB* locus SNPs which were significant for productive TCRs and not significant for non-productive TCRs occurred for the usage of genes which have both productive and non-productive alleles [125]. Only SNP associations which were significant for either productive TCRs, non-productive TCRs, or both are shown here. There were 15 significant associations which were not located within the MHC, *TRB* or ZNF443/ZNF709 loci and are not shown here. The solid black horizontal and vertical lines correspond to the genome-wide Bonferroni-corrected P-value significance threshold of 5.09×10^{-11}. The dashed black line represents the non-productive -log10(P-value) equals productive -log10(P-value) line.

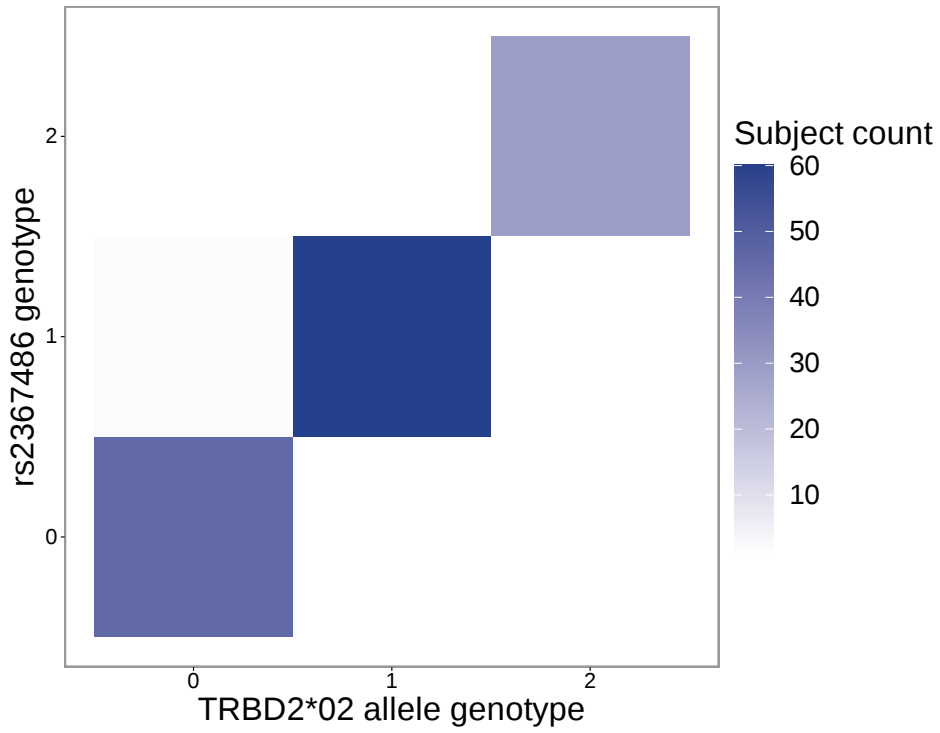

Figure A.4: The SNP genotype for the SNP (rs2367486) most significantly associated with 5' end D-gene trimming within the *TRB* locus is also associated with *TRBD2*02* allele genotype. Specifically, SNP genotype and *TRBD2*02* allele genotype are significantly correlated ($P < 2.2 \times 10^{-16}$ and $\chi^2 = 259.3$) using a chi-square test of independence. The Y-axis integer genotypes correspond to the number of minor alleles within the rs2367486 SNP genotype. The X-axis integer genotypes correspond to the number of *TRBD2*02* alleles within the *TRBD2* gene locus genotype.

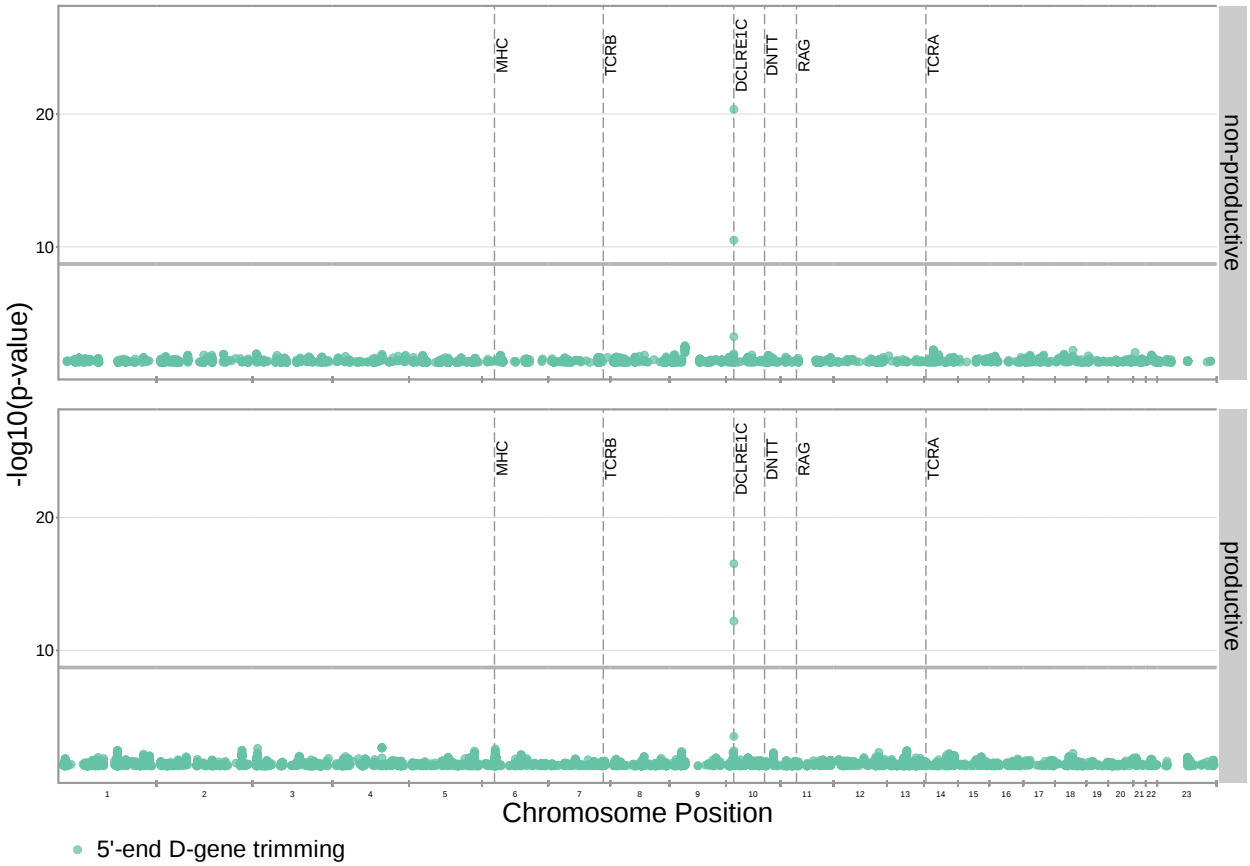

Figure A.5: Significant associations are no longer observed between 5' end D-gene trimming and variation in the *TRB* locus after correcting for *TRBD2* allele genotype in our model formulation. Further, four new significant associations are present between 5' end D-gene trimming and variation in the *DCLRE1C* locus. Only SNP associations whose $P < 5 \times 10^{-2}$ are shown here. All genome-wide 3' end D-gene trimming associations fell above this plotting threshold. The gray horizontal line corresponds to a P-value of 1.94×10^{-9} (calculated using whole-genome Bonferroni correction, see Methods).

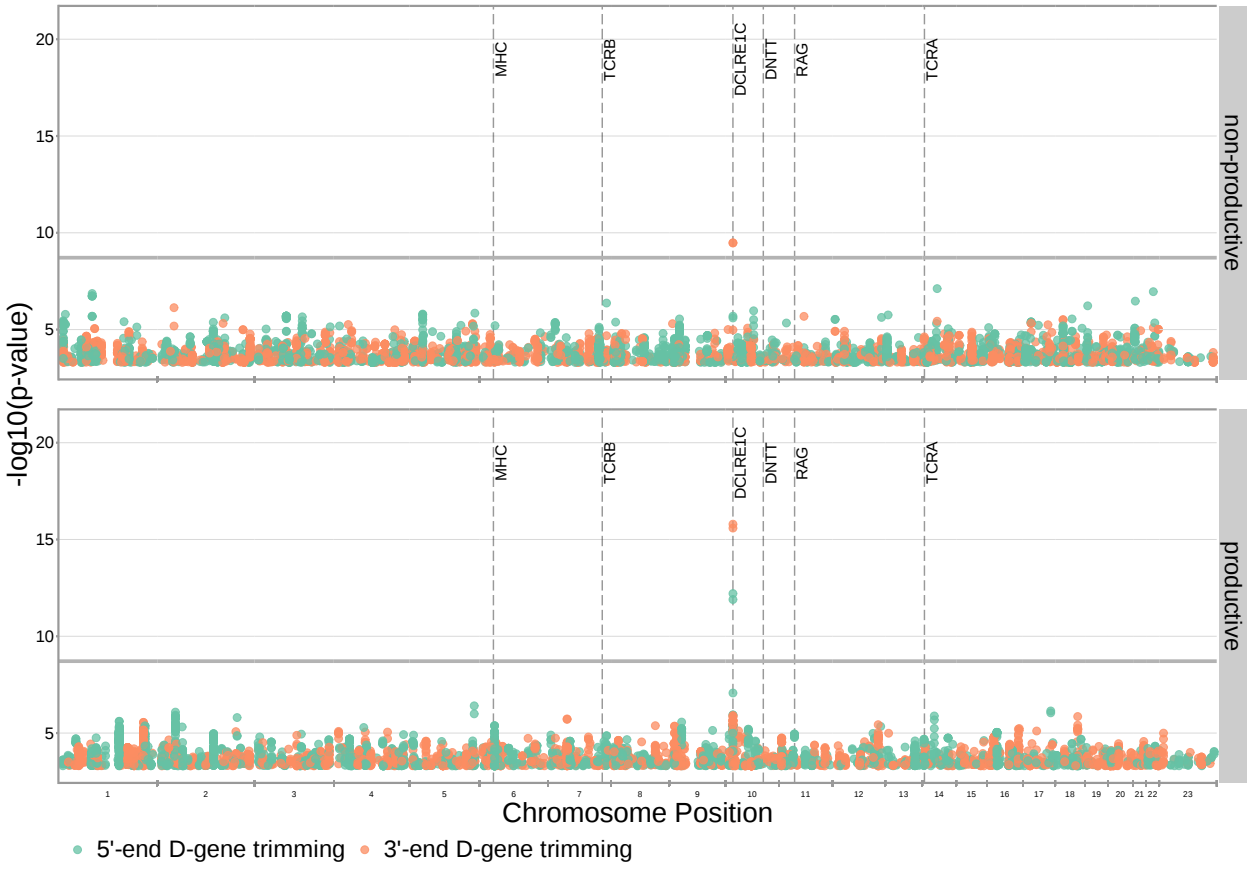

Figure A.6: Significant associations are also no longer observed between 5' end D-gene trimming and variation in the *TRB* locus when restricting the analysis to TCRs which contain *TRBJ1* genes (and consequently contain *TRBD1*). Additionally, two new associations are present between 5' end D-gene trimming and variation in the *DCLRE1C* locus for productive TCRs. Four new associations are present between 3' end D-gene trimming and variation in the *DCLRE1C* locus. Only SNP associations whose $P < 5 \times 10^{-4}$ are shown here. The gray horizontal line corresponds to a P-value of 1.94×10^{-9} (calculated using whole-genome Bonferroni correction, see Methods).

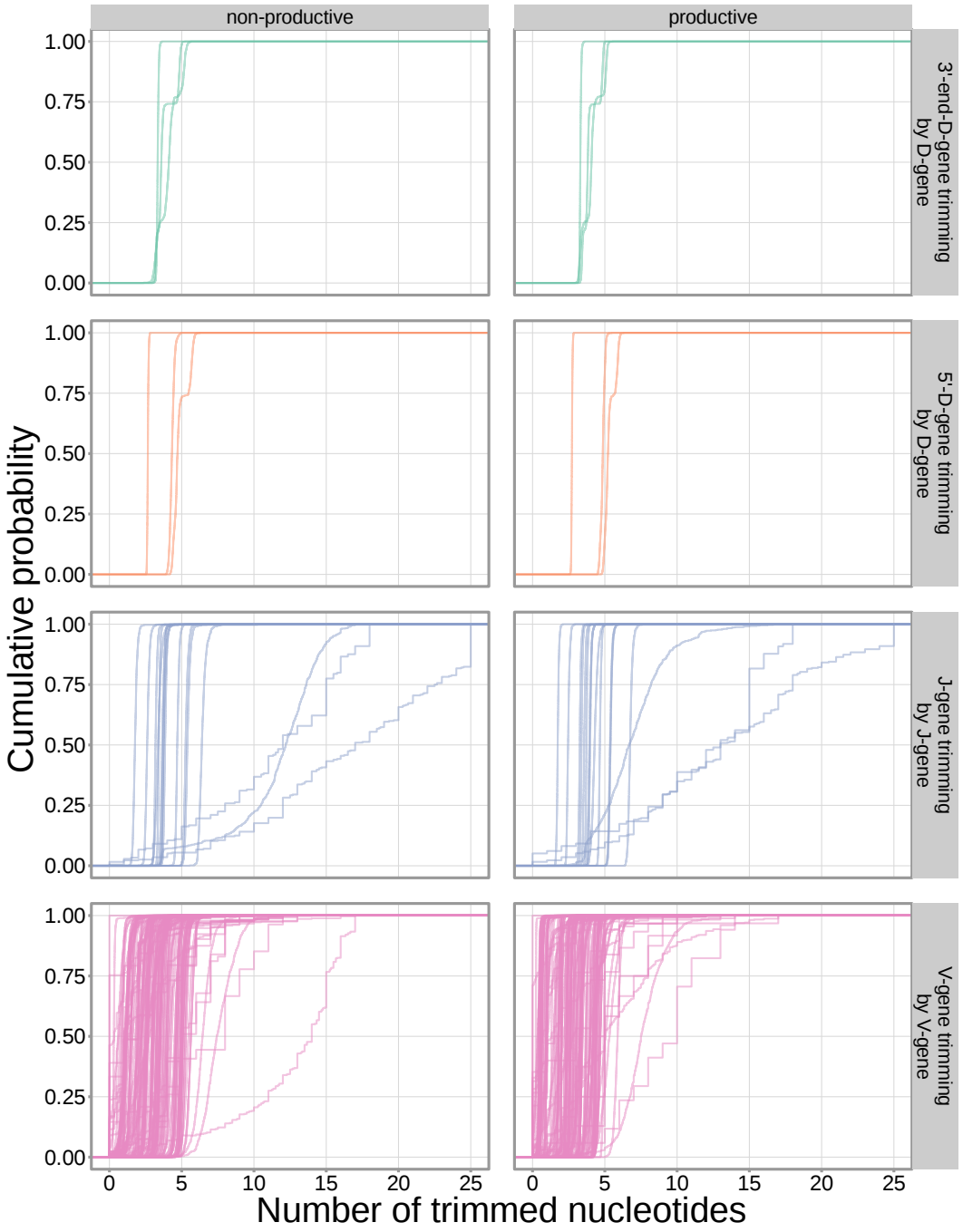

Figure A.7: The extent of nucleotide deletion varies by the gene allele identity for all gene types. An empirical cumulative distribution is drawn for each gene allele type within each indicated gene type (i.e. V-gene, D-gene, J-gene).

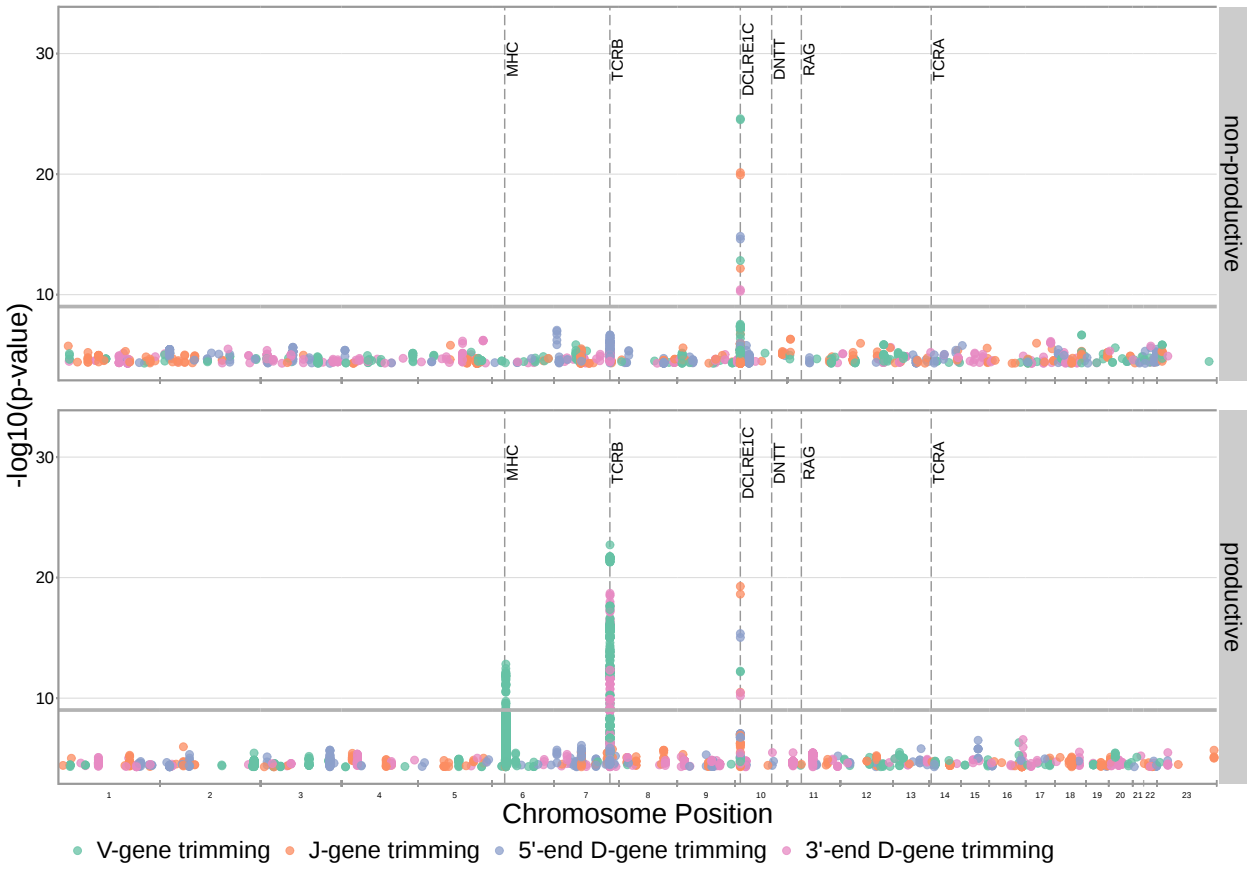

Figure A.8: Significant SNP associations are located within the MHC, *TRB* and *DCLRE1C* loci for all four trimming types when calculating the strength of association without conditioning out effects mediated by gene choice. Earlier findings relating variations in MHC and *TRB* to gene usage changes, however, indicate that many of these associations are likely artefactual. Only SNP associations whose $P < 5 \times 10^{-5}$ are shown here. The gray horizontal line corresponds to a P-value of 9.68×10^{-10}.

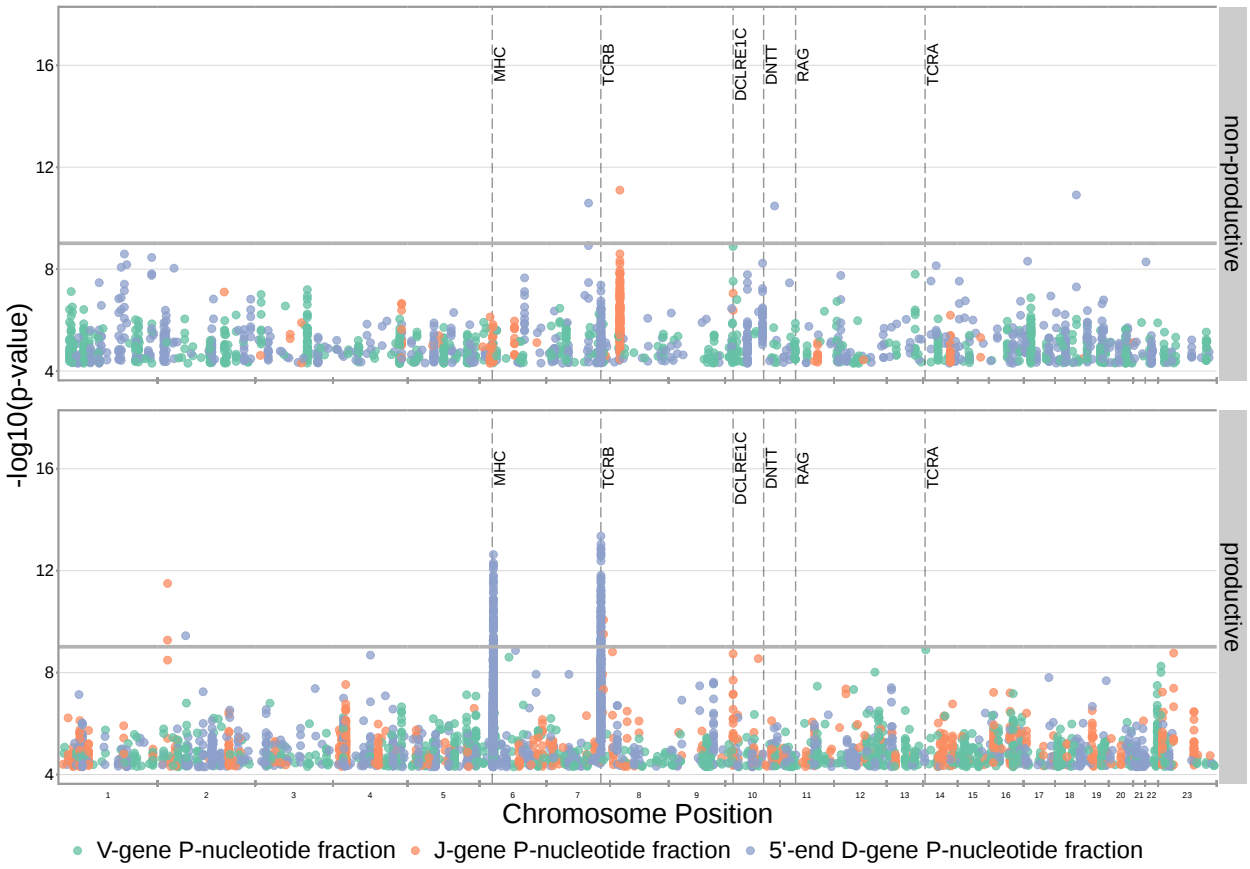

Figure A.9: The fraction of TCRs containing P-nucleotides is not significantly associated with *DCLRE1C* locus SNPs. However, significant associations are present within the *TRB* and MHC loci for the fraction of non-D-gene-trimmed, productive TCRs containing 5' end D-gene P-nucleotides. Only SNP associations whose $P < 5 \times 10^{-5}$ are shown here. The gray horizontal line corresponds to a P-value of 9.68×10^{-10} (calculated using whole-genome Bonferroni correction, see Methods).

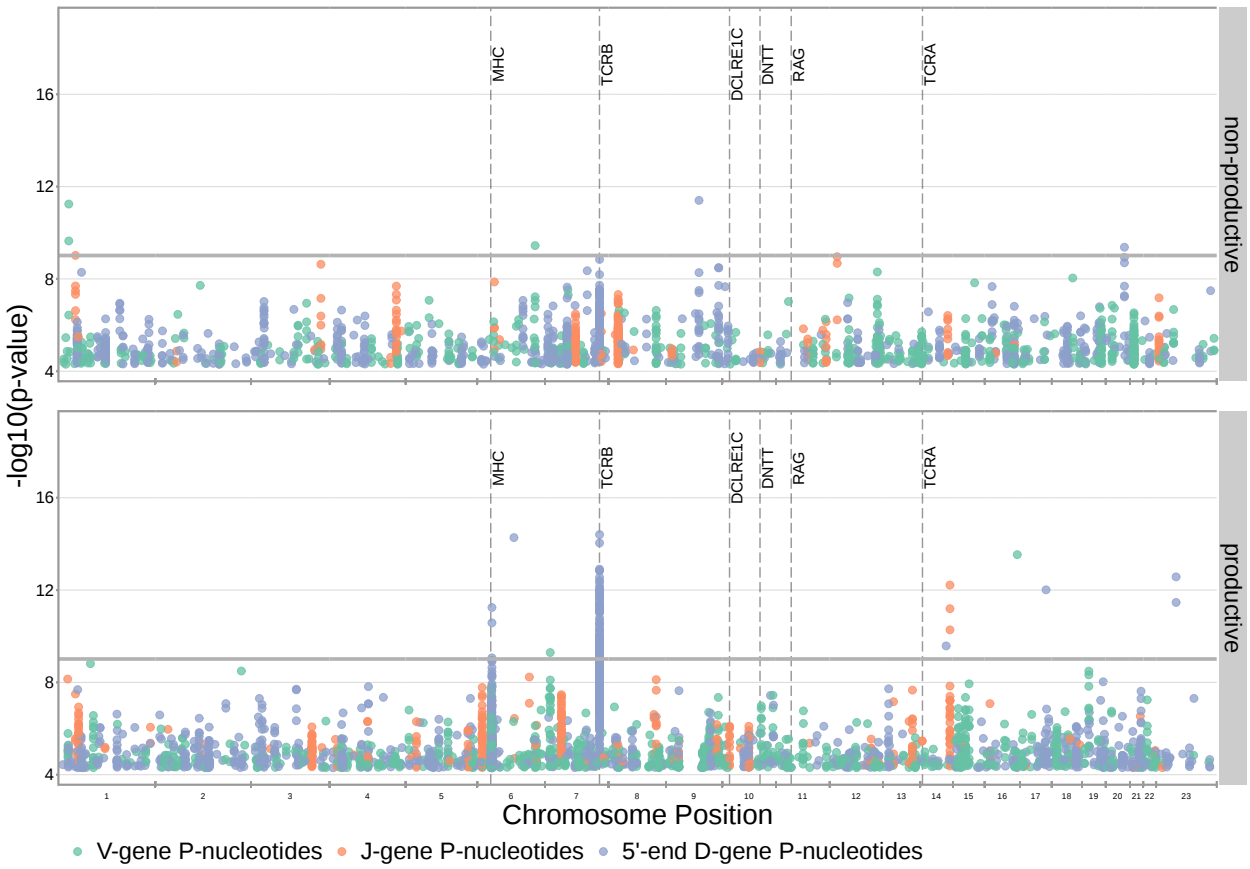

Figure A.10: The number of P-nucleotides is not significantly associated with *DCLRE1C* locus SNPs. However, significant associations are present within the *TRB* and MHC loci. Only SNP associations whose $P < 5 \times 10^{-5}$ are shown here. The gray horizontal line corresponds to a Bonferroni-corrected whole-genome P-value significance threshold of 9.68×10^{-10} (see Methods).

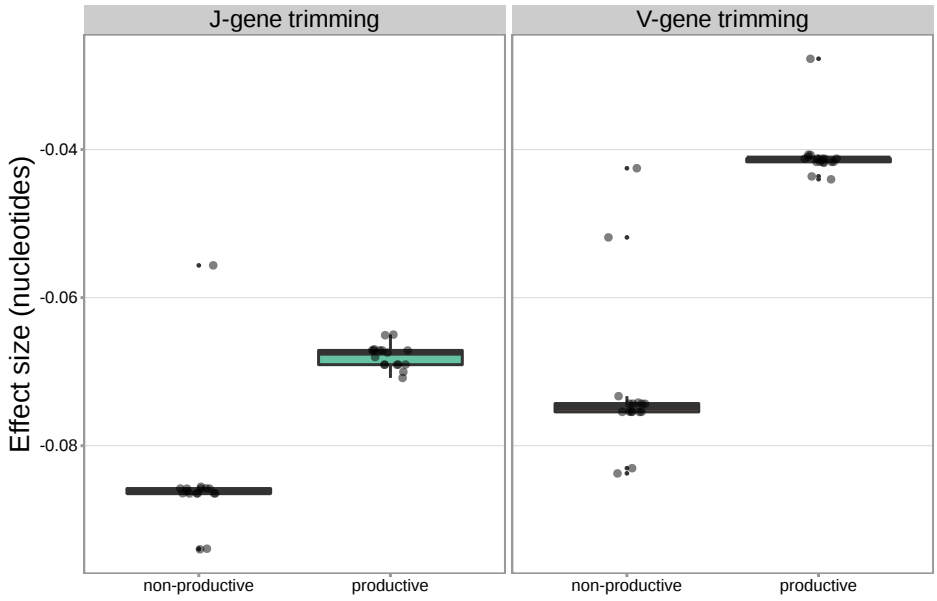

Figure A.11: For the significantly associated *DCLRE1C* locus SNPs, the magnitudes of the effects were greater for non-productive TCRs compared to productive TCRs for both V-gene trimming and J-gene trimming.

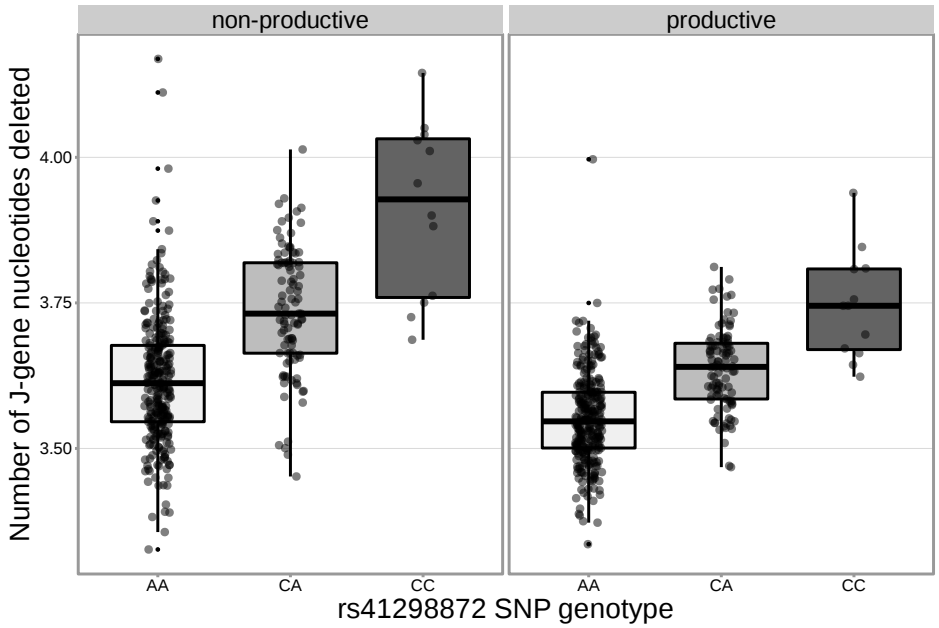

Figure A.12: The extent of J-gene trimming changes as a function of SNP genotype for the SNP (rs41298872) most significantly associated with J-gene trimming within the *DCLRE1C* locus. Only TCRs containing *TRBJ1-1*01* (the most frequently used *TRBJ1* gene across subjects) were included when calculating the average number of J-gene nucleotides deleted for each subject.

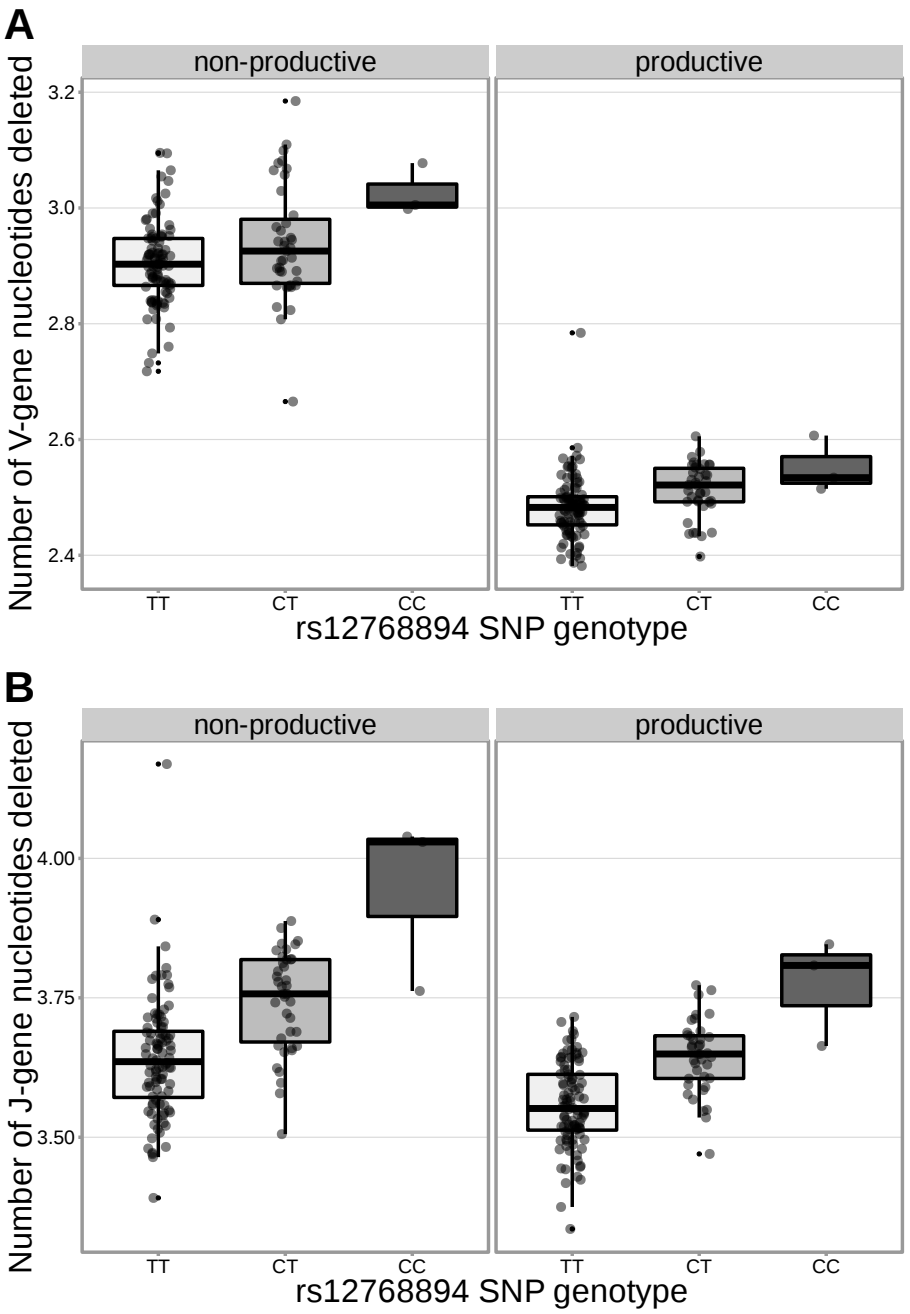

Figure A.13: The extent of V- and J-gene trimming of productive and non-productive TCRβ chains changes as a function of SNP genotype within the discovery cohort for a non-synonymous *DCLRE1C* SNP (rs12768894, c.728A>G). Only TCRs containing *TRBJ1-1*01* (the most frequently used *TRBJ1* gene across subjects) were included when calculating the average number of J-gene nucleotides deleted for each subject. Only TCRs containing *TRBV5-1*01* (the most frequently used *TRBV* gene across subjects) were included when calculating the average number of V-gene nucleotides deleted for each subject.

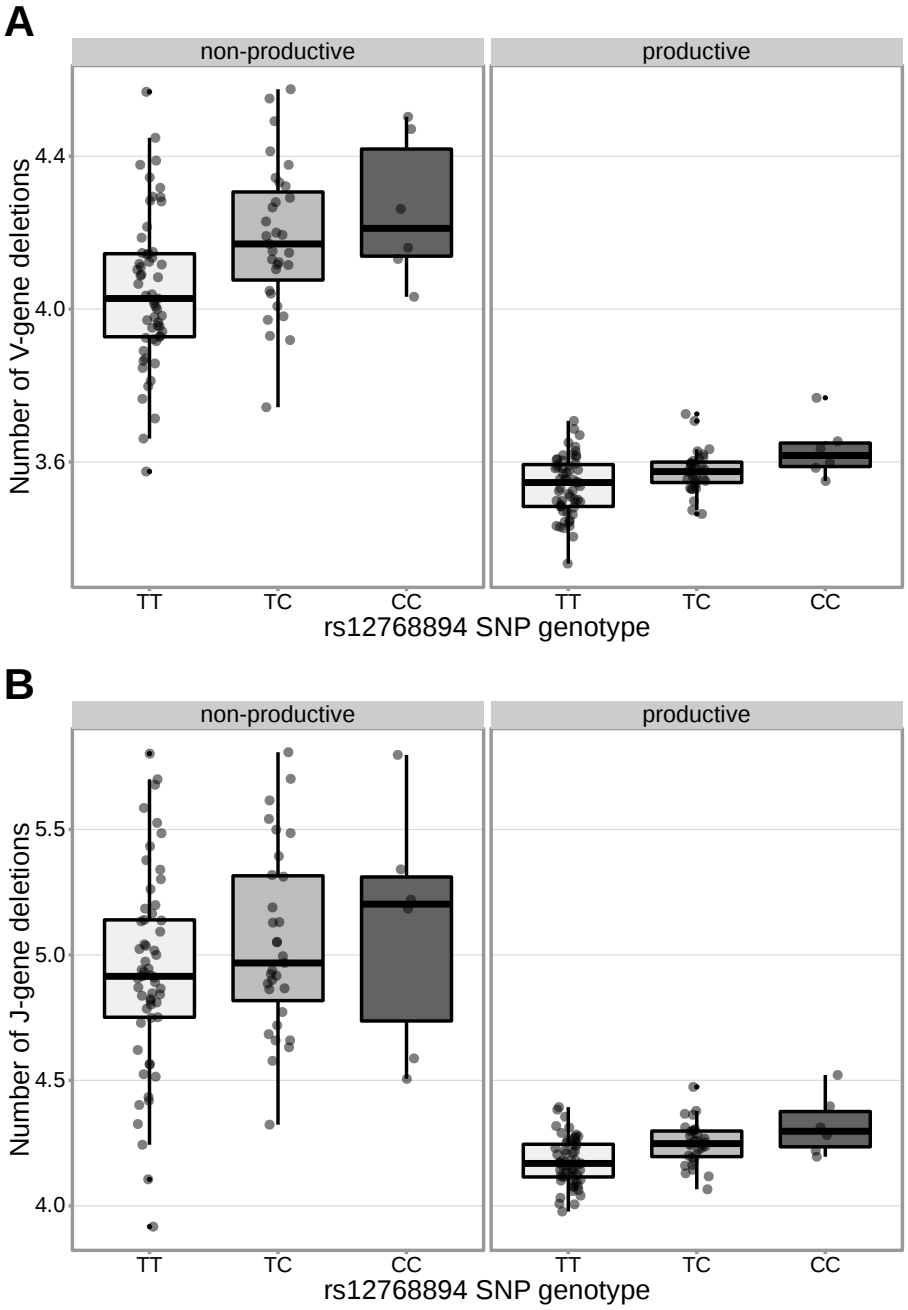

Figure A.14: The extent of V-gene trimming (**A**) of productive and non-productive TCRβ chains and J-gene trimming (**B**) of productive TCRβ chains changes as a function of SNP genotype within the validation cohort for a non-synonymous *DCLRE1C* SNP (rs12768894, c.728A>G). The average number of nucleotides deleted was calculated across all TCRβ chains for each subject, regardless of gene-usage.

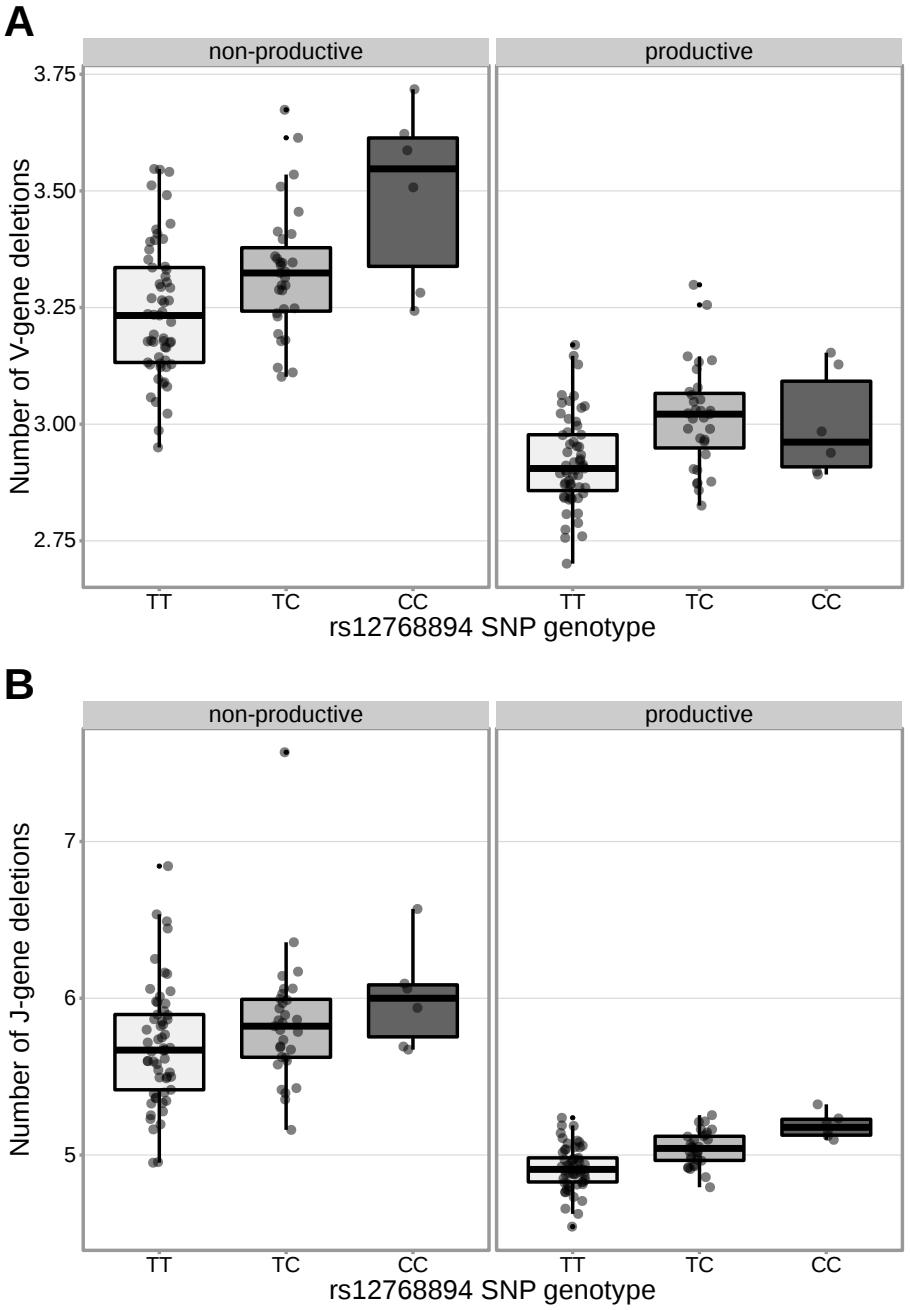

Figure A.15: The extent of V- (**A**) and J-gene (**B**) trimming of productive and non-productive TCRα chains changes as a function of SNP genotype within the validation cohort for a non-synonymous *DCLRE1C* SNP (rs12768894, c.728A>G). The average number of nucleotides deleted was calculated across all TCRα chains for each subject, regardless of gene-usage.

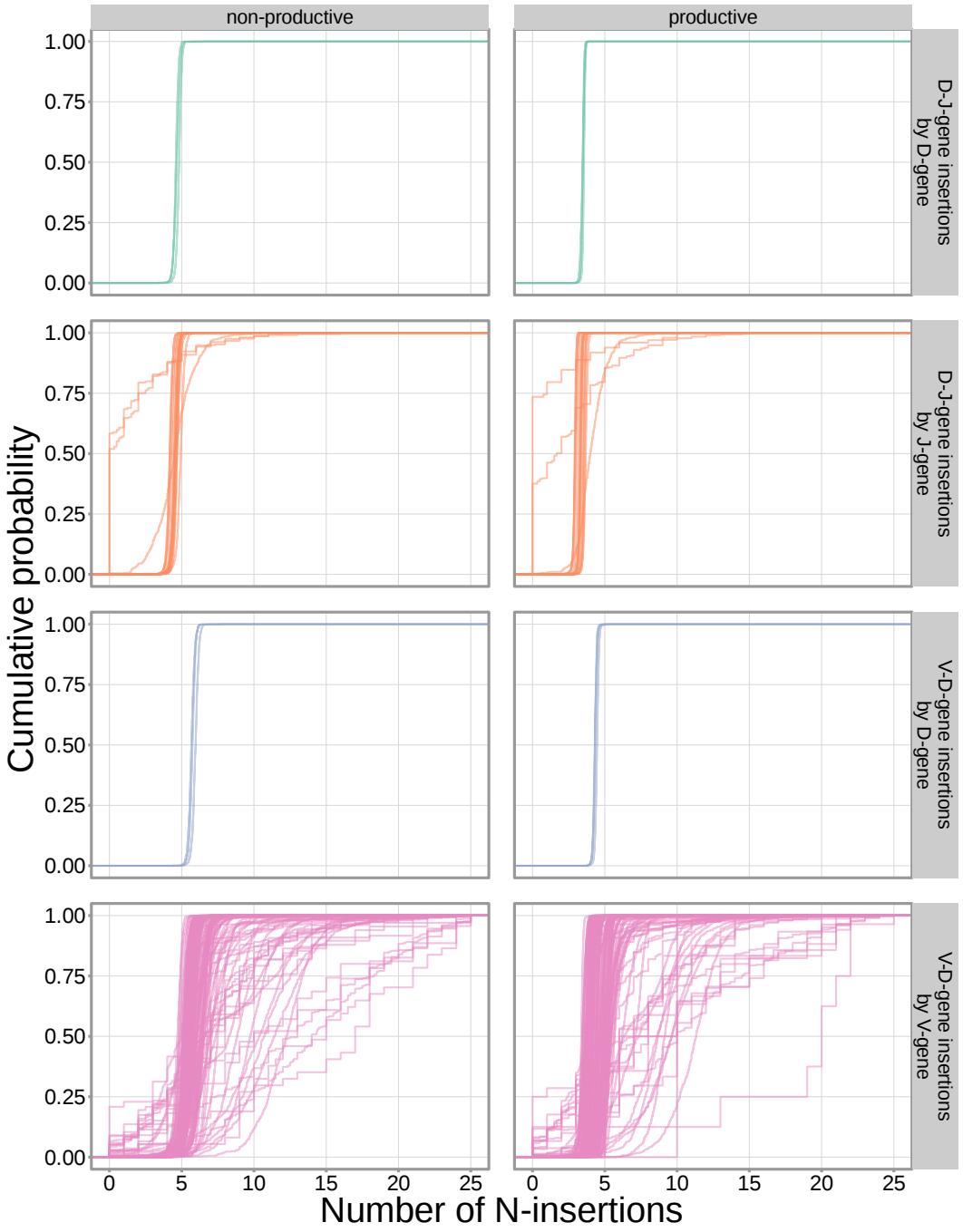

Figure A.16: The extent of N-insertion does not vary substantially by the gene allele identity for any gene type. An empirical cumulative distribution is drawn for each gene allele type within each indicated gene type (i.e. V-gene, D-gene, J-gene).

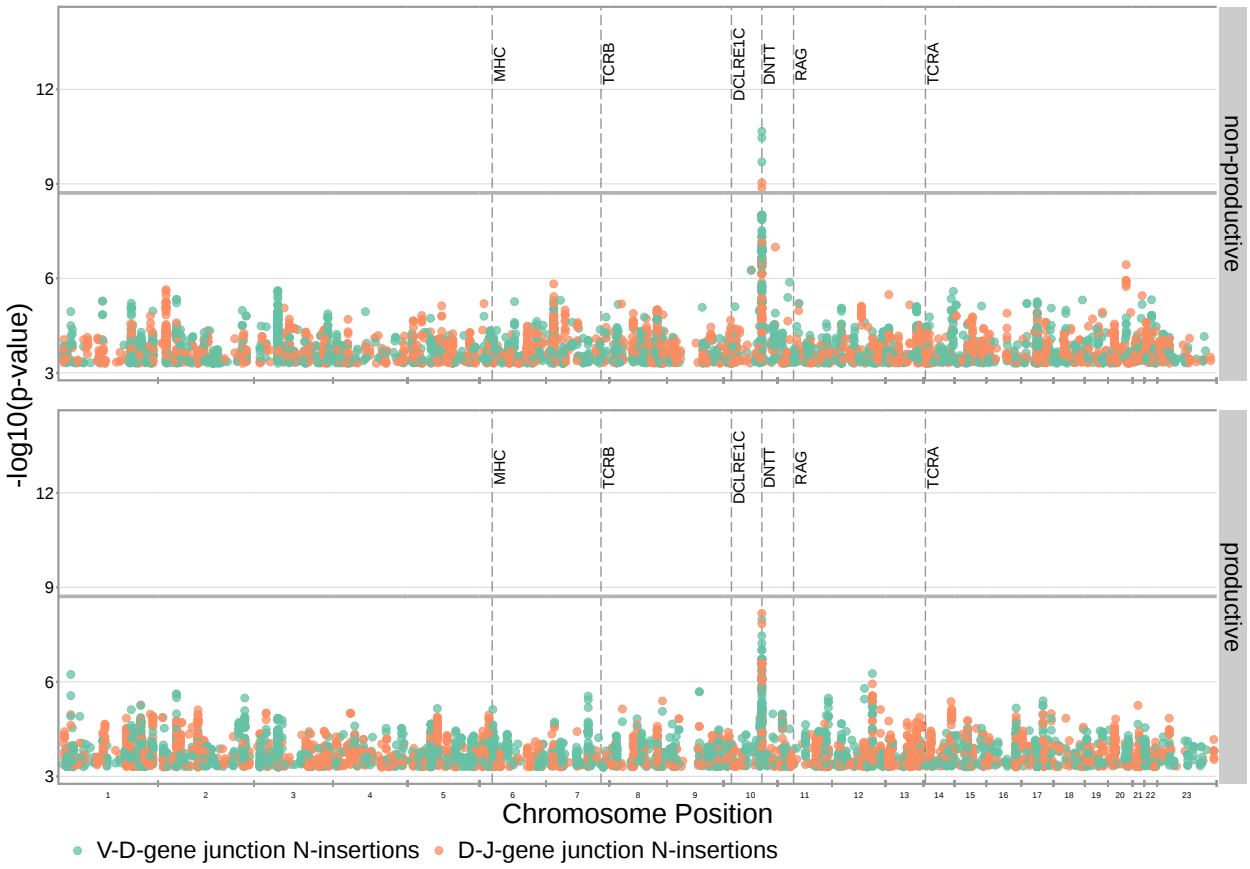

Figure A.17: Significant associations continue to be observed within the *DNTT* locus for both V-D- and D-J-gene-junction N-insertions when restricting the analysis to TCRs which contain *TRBJ1* genes (and consequently contain *TRBD1*. Only SNP associations whose $P < 5 \times 10^{-4}$ are shown here. The gray horizontal line corresponds to a Bonferroni-corrected P-value significance threshold of 1.94×10^{-9} (calculated using whole-genome Bonferroni correction, see Methods).

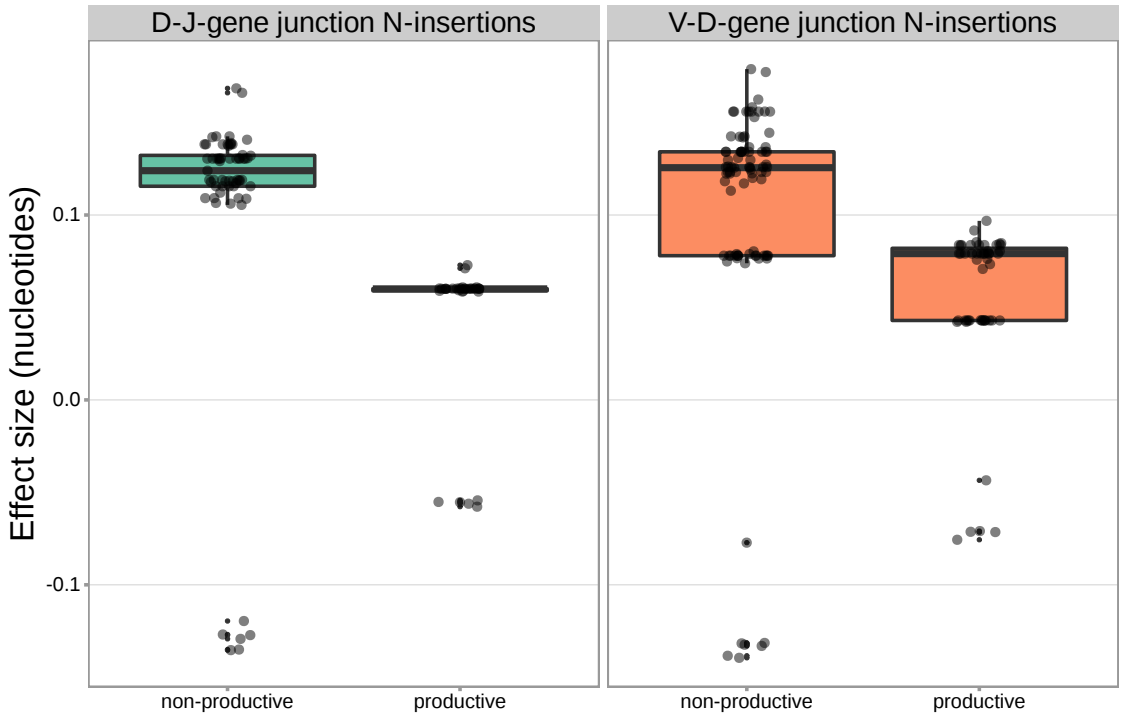

Figure A.18: For these significant *DNTT* locus SNP associations, the magnitudes of the effects were greater for non-productive TCRs compared to productive TCRs for both V-D-gene junction N-insertion and D-J-gene junction N-insertion.

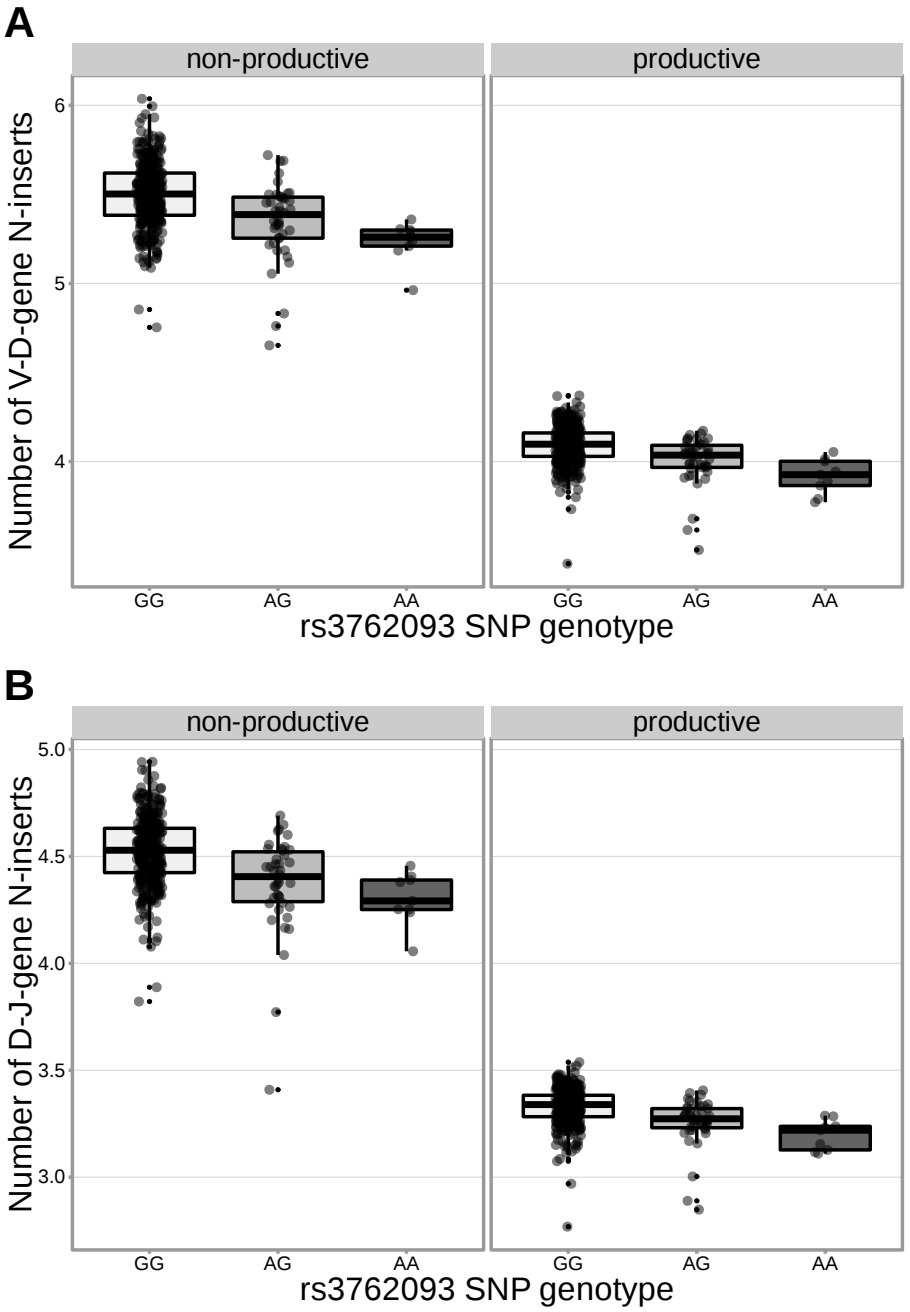

Figure A.19: The extent of V-D and D-J N-insertion of productive and non-productive TCRβ chains changes as a function of SNP genotype within the discovery cohort for an intronic *DNTT* SNP (rs3762093). The average number of N-insertions was calculated across all TCRβ chains for each subject.

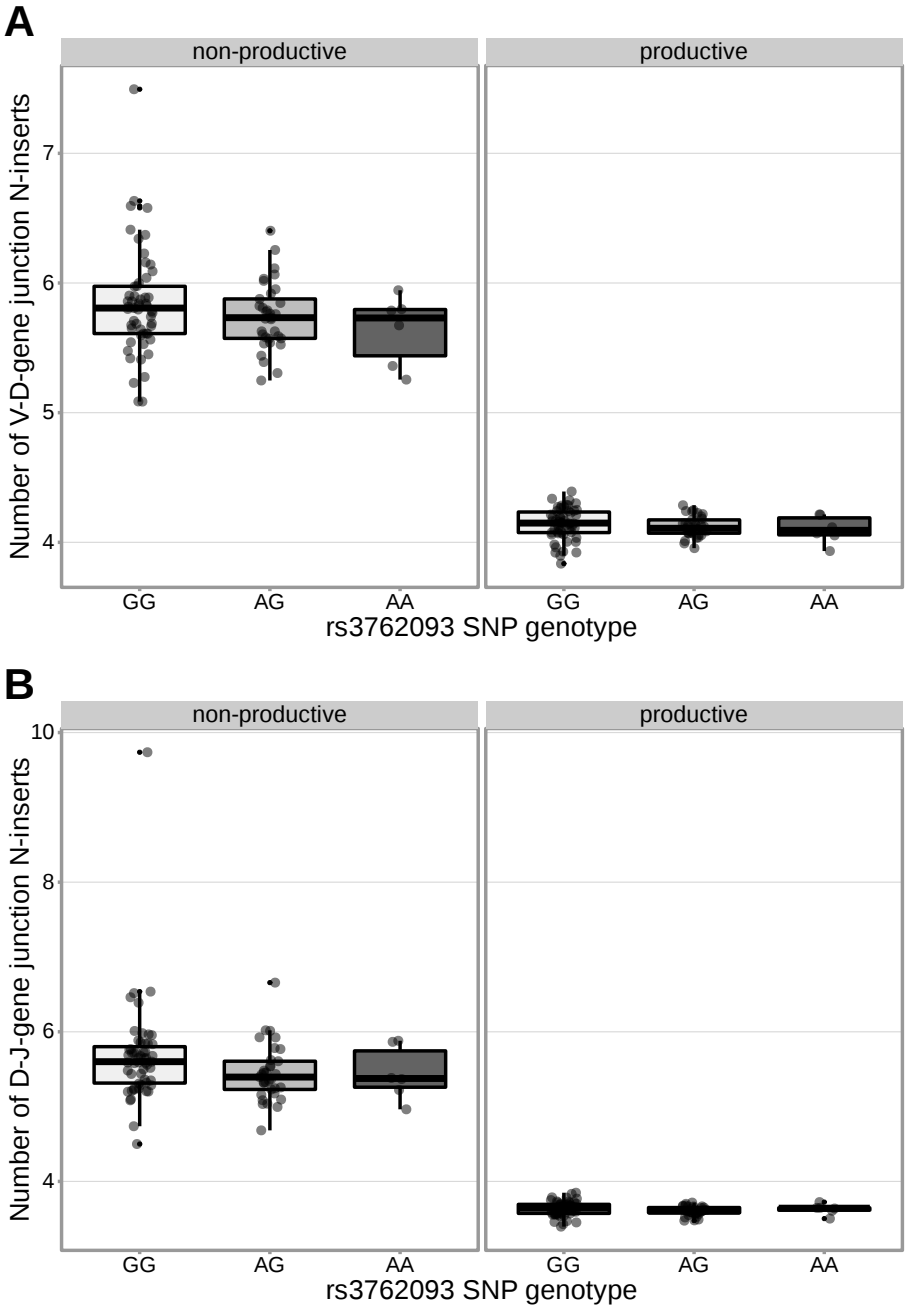

Figure A.20: An intronic SNP (rs3762093) within the *DNTT* gene locus is not strongly associated with the number of V-D (**A**) or D-J (**B**) N-inserts within productive or non-productive TCRβ chains in the validation cohort. However, the direction of the effect is the same as the discovery cohort for all N-insertion and productivity types. The average number of N-insertions was calculated across all TCRβ chains for each subject.

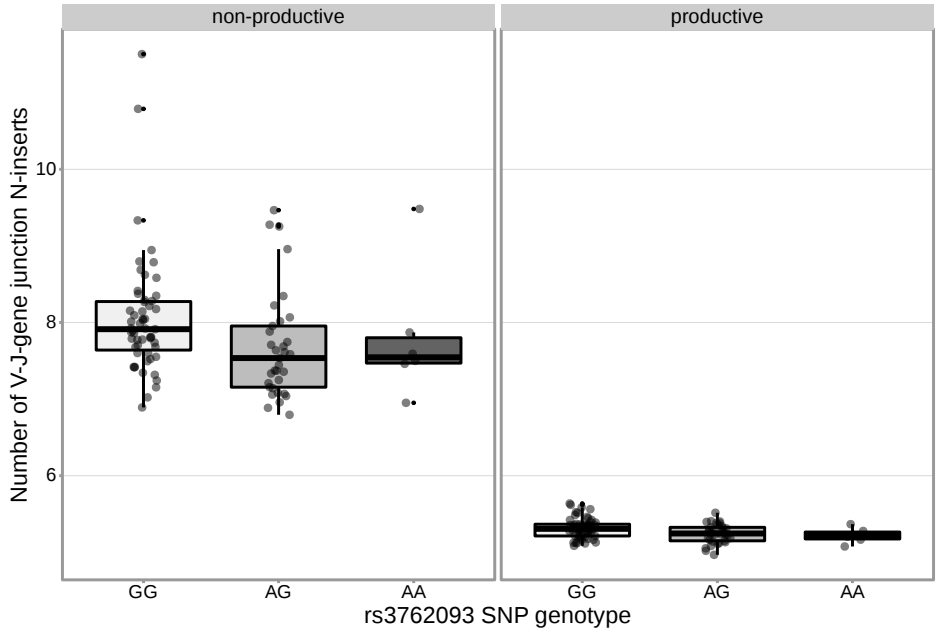

Figure A.21: An intronic SNP (rs3762093) within the *DNTT* gene locus is significantly associated with the number of V-J N-inserts for productive TCRα chains in the validation cohort. This SNP is not significantly associated with the number of V-J N-inserts for non-productive TCRα chains in the validation cohort. The average number of N-insertions was calculated across all TCRα chains for each subject.

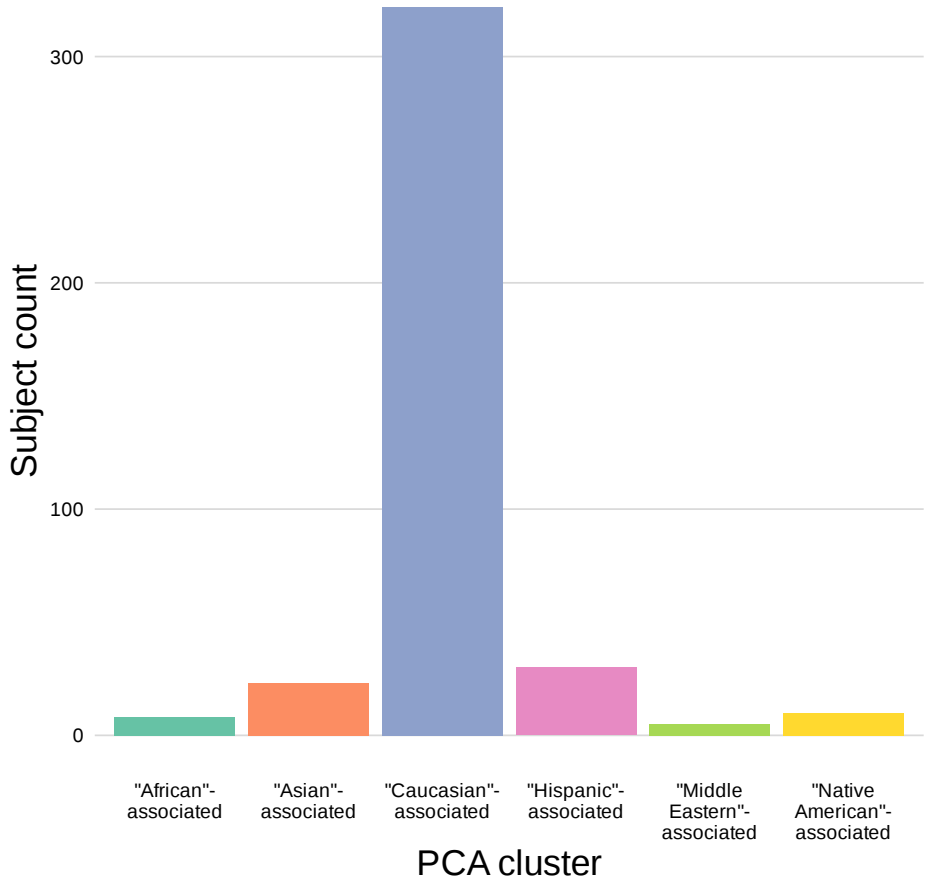

Figure A.22: The discovery cohort population mean is dominated by subjects in the “Caucasian”-associated PCA-cluster. Of the 398 subjects in the sample population, 81% are in the “Caucasian”-associated PCA-cluster.

Table A.1: Key resources table.

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Software, Algorithm	TCRdist	[67]		Version 0.0.2; Software can be found on GitHub (`https://github.com/phbradley/tcr-dist`)
Software, Algorithm	migec	[70]	RRID: SCR_016337	Version 1.2.9; Software can be found on GitHub (`https://github.com/mikessh/migec`)
Antibody	CD3-PerCP eFluor710 (Mouse monoclonal)	Thermo Fisher Scientific	Cat: 46-0037-42; RRID: AB_1834395	0.012 μg per 1 million cells (1:100)
Antibody	CD4-BV650 (Mouse monoclonal)	BD Biosciences	Cat: 563875; RRID: AB_2687486	2 μl per 1 million cells (1:50)
Antibody	CD8-APC Fire750 (Mouse monoclonal)	Biolegend	Cat: 344746; RRID: AB_2572095	0.1 μg per 1 million cells (1:100)
Antibody	TCRγ/δ-PE Cy7 (Mouse monoclonal)	Biolegend	Cat: 331222; RRID: AB_2562891	1 μg per 1 million cells (1:40)
Other	Fc Block	BD Biosciences	Cat: 564220; RRID: AB_2728082	2.5 μg per 1 million cells (1:20)
Other	Live/Dead Aqua	Tonbo Biosciences	Cat: 13-0870-T100	1 μl per 1 million cells (1:100)
Commercial assay, kit	Qiagen QIAamp DNA Mini Kit	Qiagen	Cat: 51306	
Commercial assay, kit	Taqman SNP Genotyping Assay	Thermo Fisher Scientific	Cat: 4351379	
Commercial assay, kit	TaqMan Genotyping Master Mix	Thermo Fisher Scientific	Cat: 4371353	

Appendix B

SUPPLEMENTARY METHODS FOR CHAPTER 2

B.1 Correcting for TRBD2-allele-SNP genotype linkage in TCR feature associations with the TRB locus SNPs

Within the *TRB* locus, we noted that SNP genotypes were associated with *TRBD2* allele genotype (Figure A.4). Associations between gene-alleles and *TRB* locus SNP genotypes, if present, may produce false positive associations when implementing the "gene-conditioned model" to infer associations between SNPs and TCR repertoire features, conditional on gene usage. To explore this phenomenon further, we zoomed in to the *TRB* locus and incorporated a *TRBD2* allele genotype correction procedure into our model formulation. As such, to quantify the association strength between each *TRB* locus SNP and TCR feature, conditional on gene usage and correcting for population-substructure-related effects as in (2.7), while incorporating *TRBD2* allele genotype correction terms, we fit the model:

$$\bar{y}_{im} = z_i \cdot \alpha_j + x_{ij} \cdot \beta_{1j} + \beta_0 + \gamma_{jm} + \sum_{p=1}^{8} \beta_{2jp} \cdot P_{ip} + \epsilon_{ijm}$$

where z_i represents the qualitative *TRBD2* allele genotype status for subject i such that $z_i \in$ {"TRBD2*01 homozygous", "heterozygous", "TRBD2*02 homozygous"}, α_j is the *TRBD2* allele genotype effect for SNP j, and the remaining variables are defined as in (2.7). With this model formulation, we can estimate each regression coefficient by solving the following

weighted least squares problem for each *TRB* SNP, TCR feature, and productivity status combination:

$$(\hat{\alpha}_j, \hat{\beta}_0, \hat{\beta}_{1j}, \hat{\gamma}_j, \vec{\hat{\beta}}_{2j}) = \underset{\alpha_j, \beta_0, \beta_{1j}, \gamma_j, \vec{\beta}_{2j}}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{m=1}^{M_t} W_{im} \cdot \left(\bar{y}_{im} - (\alpha_j z_i + \beta_0 + \gamma_{jm} + \beta_{1j} x_{ij} + \sum_{p=1}^{8} \beta_{2jp} \cdot P_{ip})\right)^2.$$

With these estimates for the *TRBD2* allele genotype and population-substructure-corrected j-th SNP effect on the amount of the TCR feature of interest, $\hat{\beta}_{1j}$, we calculated a P-value using the methods described in the methods section for the "gene-conditioned model".

B.2 Genomic inflation factor calculations

We defined the genomic inflation factor λ to be the ratio of the median of the empirically observed squared test statistic to the expected median [126–128]. For each GWAS analysis implemented using the "simple model", we used the test statistic T_j given by (2.3) for each SNP $j = \{1 \dots J\}$ tested genome-wide. For each GWAS analysis implemented using the "gene-conditioned model", it was not computational feasible to calculate a test statistic T_j for all SNPs tested genome-wide using the bootstrapping protocol described in the "gene-conditioned model" methods section. Thus, instead, we randomly sampled 10,000 SNPs and calculated the test statistic T_j for each SNP in the random subset. Let $S = \{T_1^2, ..., T_J^2\}$ be the set of all squared test statistics. As such,

$$\lambda = \frac{\text{median}(S)}{0.456}$$

where 0.456 is the median of a chi-squared distribution with one degree of freedom. If the GWAS analysis results follow the chi-squared distribution, the expected value of λ is 1. Thus, when $\lambda < 1.03$, we concluded that there was no evidence of systemic population-

substructure-related bias in the analysis [72, 128].

B.3 Conditional analysis to test for multiple independent association signals

Within the *DNTT* and *DCLRE1C* loci, we performed a stepwise series of nested regression analyses to test for independent SNP associations within each locus for N-insertion and nucleotide trimming, respectively. We used the same models and covariates as the primary analyses ("simple model" for associations between N-insertion and *DNTT* variation and the "gene-conditioned model" for associations between nucleotide trimming and *DCLRE1C* variation) and included the most significant SNP within each locus as an additional covariate. We inferred the association between each SNP within each locus and the TCR feature of interest using this new conditional model and considered significant associations at a gene-level Bonferroni-corrected significance threshold for each locus. From here, we repeated this analysis (if necessary), identifying and adding additional SNPs one-by-one as a covariate to each successive model. Once the P-value of top SNP within the locus was no longer significant, we concluded the analysis. SNPs which were added as as additional covariates in the final conditional model were considered to be independent signals.

B.4 Quantifying TRBD2 allele associations with the TRB locus SNPs

For each significantly associated SNP within the *TRB* locus as shown in Figure 2.3, we compared SNP genotype to *TRBD2* allele genotype across all subjects. We used Pearson correlation to measure the correlation between the two genotypes.

B.5 Exploring TCR repertoire features and SNP minor allele frequency by ancestry PCA cluster

To quantify PCA cluster variation of TCR repertoire features (such as total N-insertions (V-D N-insertion and D-J N-insertion)), we first calculated an average of each TCR repertoire feature by subject and productivity status. We also calculated a population mean of each TCR repertoire feature by productivity status. Each subject was classified into one of six PCA clusters. Thus, we compared the mean of the TCR repertoire features within each PCA cluster to the population mean using a one-sample t-test to compute each P-value. We used Bonferroni multiple testing correction to adjust each P-value.

We also calculated SNP minor allele frequencies for the whole population and for each PCA cluster individually such that

$$\text{MAF}_{jr} = \frac{\sum_{i=1}^{I_r} x_{ij}}{2 * I_r}. \tag{B.1}$$

Here, MAF_{jr} is the minor allele frequency for SNP marker j and PCA cluster r, I_r is the number of individuals in the PCA cluster r, and x_{ij} is the number of alleles in the genotype of SNP marker j for subject $i \in \{1, \ldots, I_r\}$. For each SNP j, the minor allele was defined as the allele with the lowest frequency in the total population. To quantify minor allele frequency differences by PCA cluster for select SNPs within various loci of interest (i.e. *DNTT* gene), we compared the minor allele frequencies calculated within PCA-clusters to the minor allele frequencies calculated for the entire population using a one-sample t-test to compute each P-value. Again, we used Bonferroni multiple testing correction to adjust each P-value.

For both of these analyses, we used the `t_test` function from the `rstatix` package in R.

Appendix C

SUPPLEMENTARY TABLES AND FIGURES FOR CHAPTER 3

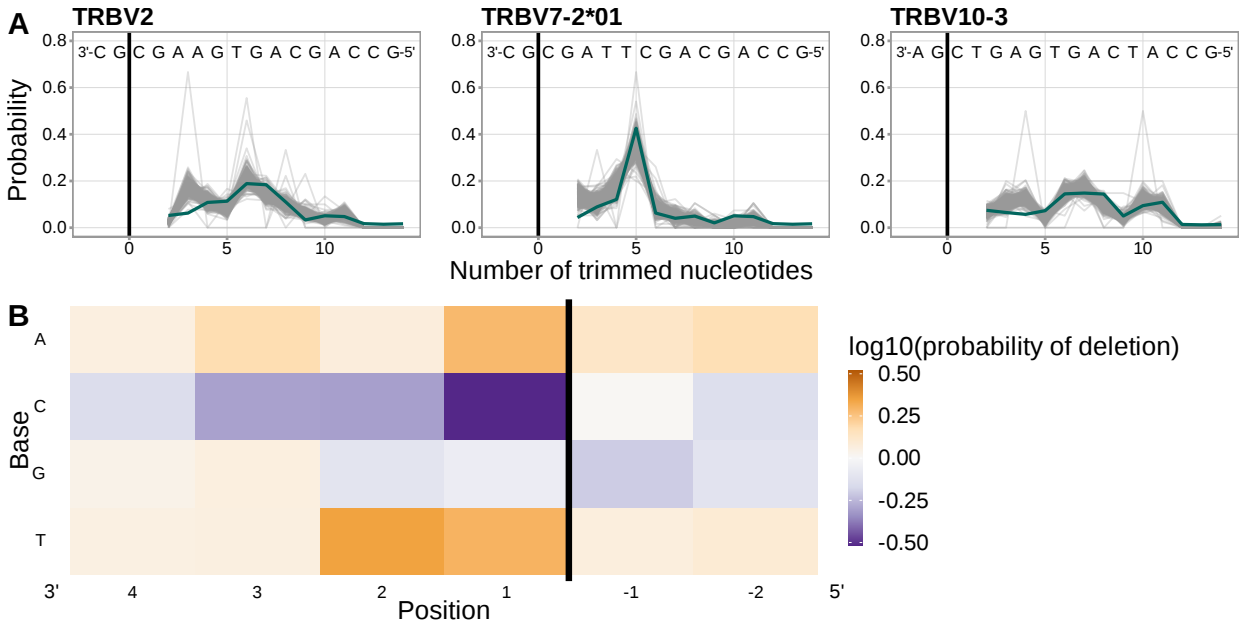

Figure C.1: Using a different, and much larger, repertoire sequencing data set, we have closely replicated previous work [15] which illustrated that a simple position-weight-matrix-style model has good predictive accuracy for many TCRβ V-genes. (**A**) Inferred trimming profiles (shown in blue) using this model have good predictive accuracy for the same V-genes highlighted in previous work (compare Figure 4A in [15]). Gene-specific trimming profiles for each individual in the training data set are shown in gray. (**B**) Position-weight-matrix of the local sequence context dependence of V-gene trimming probabilities consisting of 2 nucleotides 5' of the trimming site and 4 nucleotides 3' of the trimming site. (Note: the positions in this figure are flipped relative to the rest of the corresponding figures in this paper in order to correspond to the original figure in [15].) Positions 3' of the trimming site have a stronger effect on the probability of trimming compared to positions 5' of the trimming site. Specifically, A and T nucleotides 3' of the trimming site have a strong positive effect on the trimming probability whereas C nucleotides have a strong negative effect. The black vertical line corresponds to the trimming site. Each inferred coefficient is given as the change in log10 odds of trimming at a given site resulting from an increase in the feature value, given that all other features are held constant. The inferred coefficients show here closely resemble the previously-reported model (see Figure S11 in [15]).

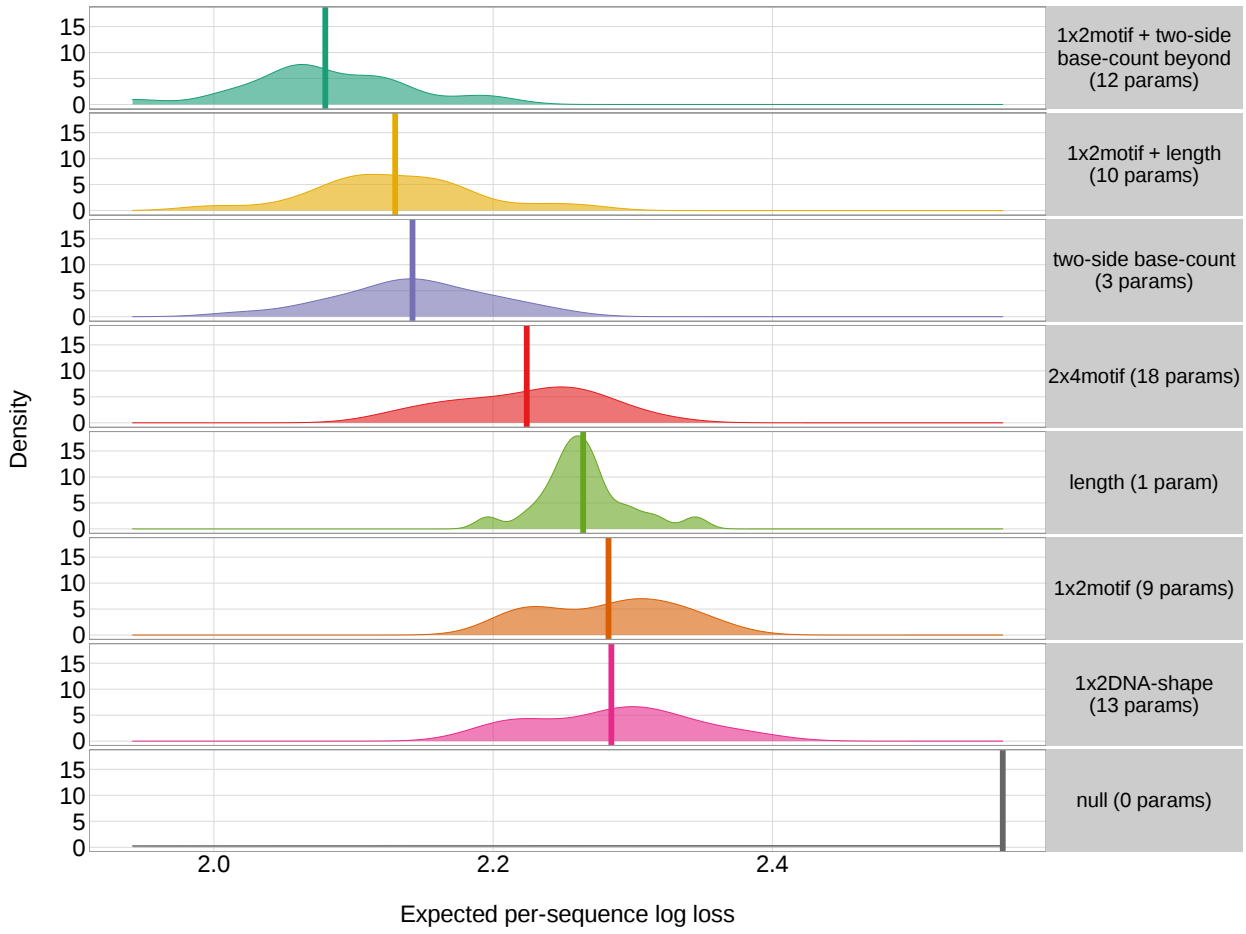

Figure C.2: For each model, there was some variation in the expected per-sequence conditional log loss values computed across the 20 random, held-out subsets of the V-gene training data set. The average expected per-sequence conditional log loss values are shown as vertical lines.

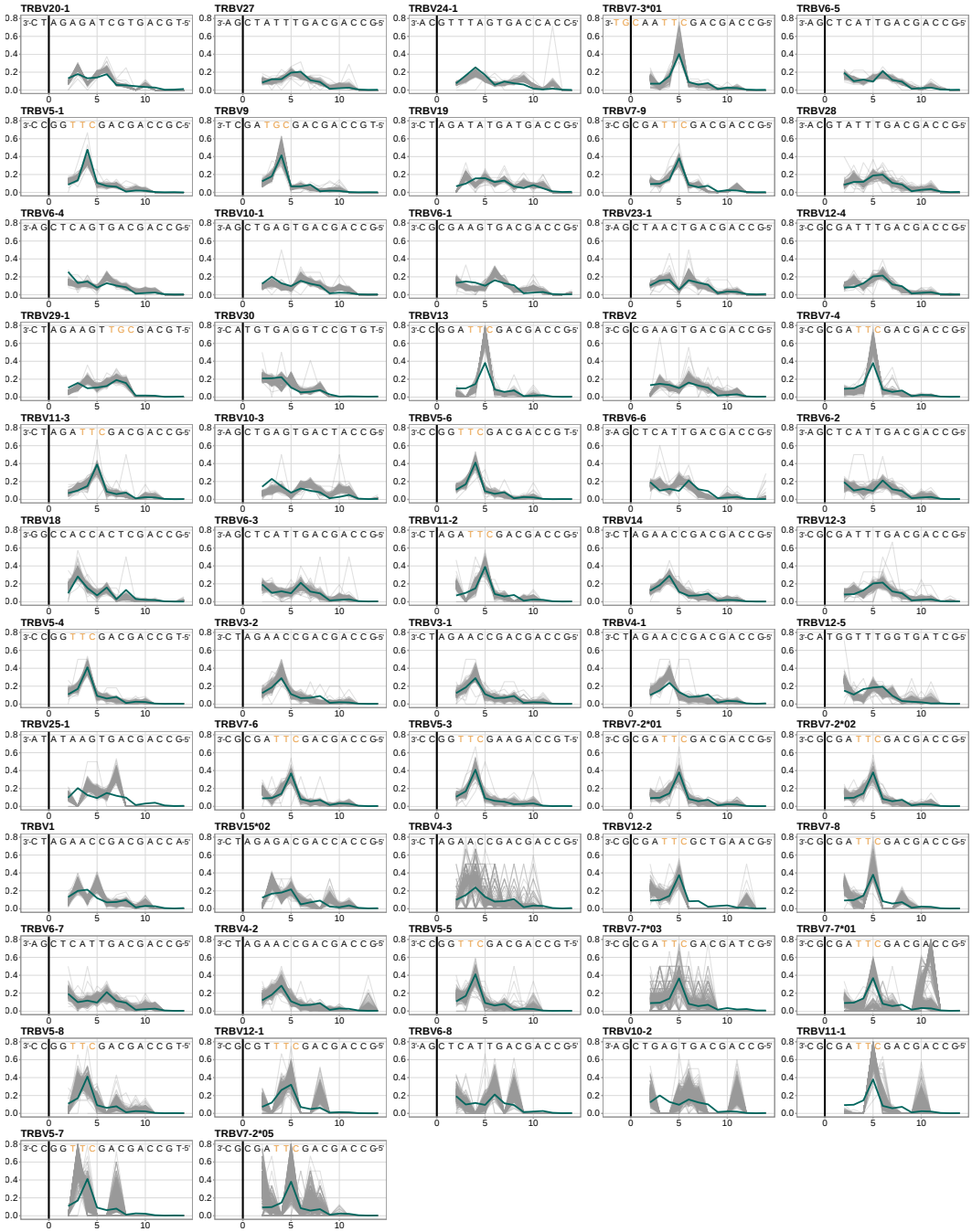

Figure C.3: Performance of the *1x2 motif + two-side base-count beyond* model across all TRB V-genes, ordered by the frequency of usage in the training data set. Inferred trimming profiles (shown in blue) using the *1x2 motif + two-side base-count beyond* model have good predictive accuracy for most V-genes. Gene-specific trimming profiles for each individual in the training data set are shown in gray. The sequence context with the highest probability of trimming (3'-TTC-5' or 3'-TGC-5' from Figure 3.4B) is highlighted in orange. The black vertical line corresponds to the trimming site.

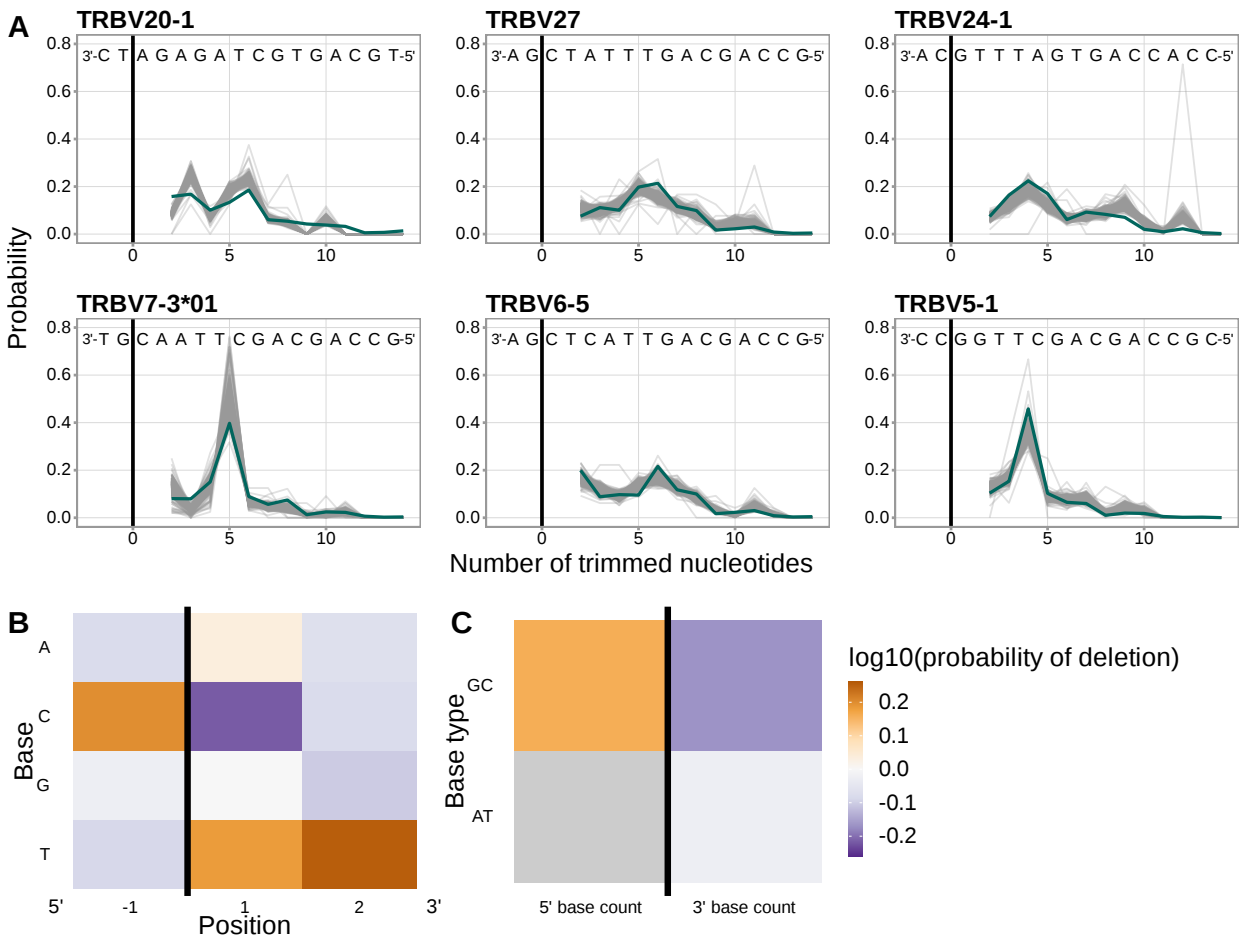

Figure C.4: (**A**) Inferred trimming profiles (shown in blue) from a *1x2 motif + two-side base-count beyond* model which includes all nucleotides 3' of the motif (regardless of double-stranded status) in the 3'-base-count term show good predictive accuracy for the most frequently used V-genes. The fit for this model is very similar to the original *1x2 motif + two-side base-count beyond* model which only includes double-stranded nucleotides in the base-count terms. Gene-specific trimming profiles for each individual in the training data set are shown in gray. (**B**) Position-weight-matrix of the local sequence context dependence of V-gene trimming probabilities consisting of 1 nucleotides 5' of the trimming site and 2 nucleotides 3' of the trimming site from fitting a *1x2 motif + two-side base-count beyond* model which uses all nucleotides 3' of the motif in the 3'-base-count term. (**C**) Inferred *two-side base-count beyond* model coefficients from fitting a *1x2 motif + two-side base-count beyond* model which uses all nucleotides 3' of the motif in the 3'-base-count term. All inferred coefficients from this model are similar to the original *1x2 motif + two-side base-count beyond* model which only includes double-stranded nucleotides in the base-count terms. The count of AT nucleotides 5' of the motif (shown in gray) was not included in this model. The black vertical line corresponds to the trimming site. Each inferred coefficient is given as the change in log10 odds of trimming at a given site resulting from an increase in the feature value, given that all other features are held constant.

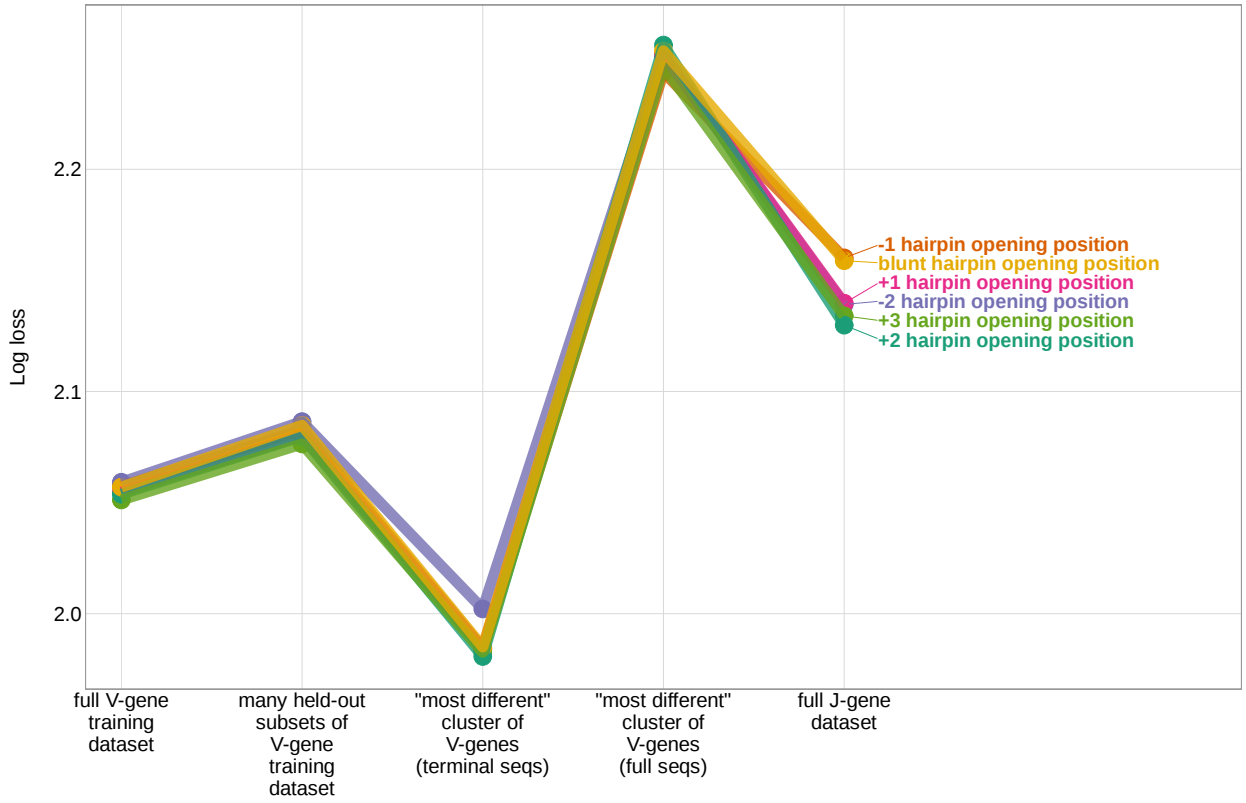

Figure C.5: Expected per-sequence conditional log loss computed for the *1x2 motif + two-side base-count beyond* model, using the full V-gene training data set, many random, held-out subsets of the V-gene training data set, a held-out subset of the V-gene training data set containing a group of V-genes defined to be the "most-different" using the terminal sequences (last 25 nucleotides of each sequence), a held-out subset of the V-gene training data set containing a group of V-genes defined to be the "most-different" using the full gene sequences, and the full J-gene data set. Each model was trained using the designated hairpin-opening-position assumption (see Appendix D for hairpin-opening-position definitions). Each model was trained using the full V-gene training data set with the held-out group or "most-different" group (if applicable) removed. Lower log loss corresponds to better a model fit.

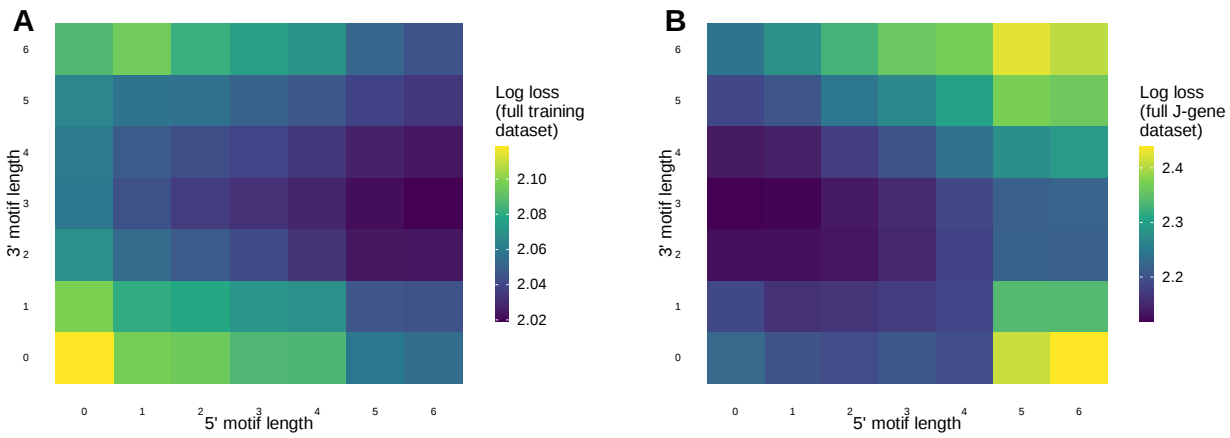

Figure C.6: (**A**) Log loss computed using the full V-gene training data set for *motif + two-side base-count beyond* models containing a varying number of bases 3' and 5' of the trimming site within the motif. (**B**) Log loss computed using the full J-gene data set for *motif + two-side base-count beyond* models containing a varying number of bases 3' and 5' of the trimming site within the motif. Each model was trained using the full V-gene training data set as described in the Materials and Methods. Lower log loss corresponds to better a model fit. Models containing small motifs have worse model fit when evaluating log loss using the full V-gene training data set (A), but have better model fit when evaluating log loss using the full J-gene data set (B).

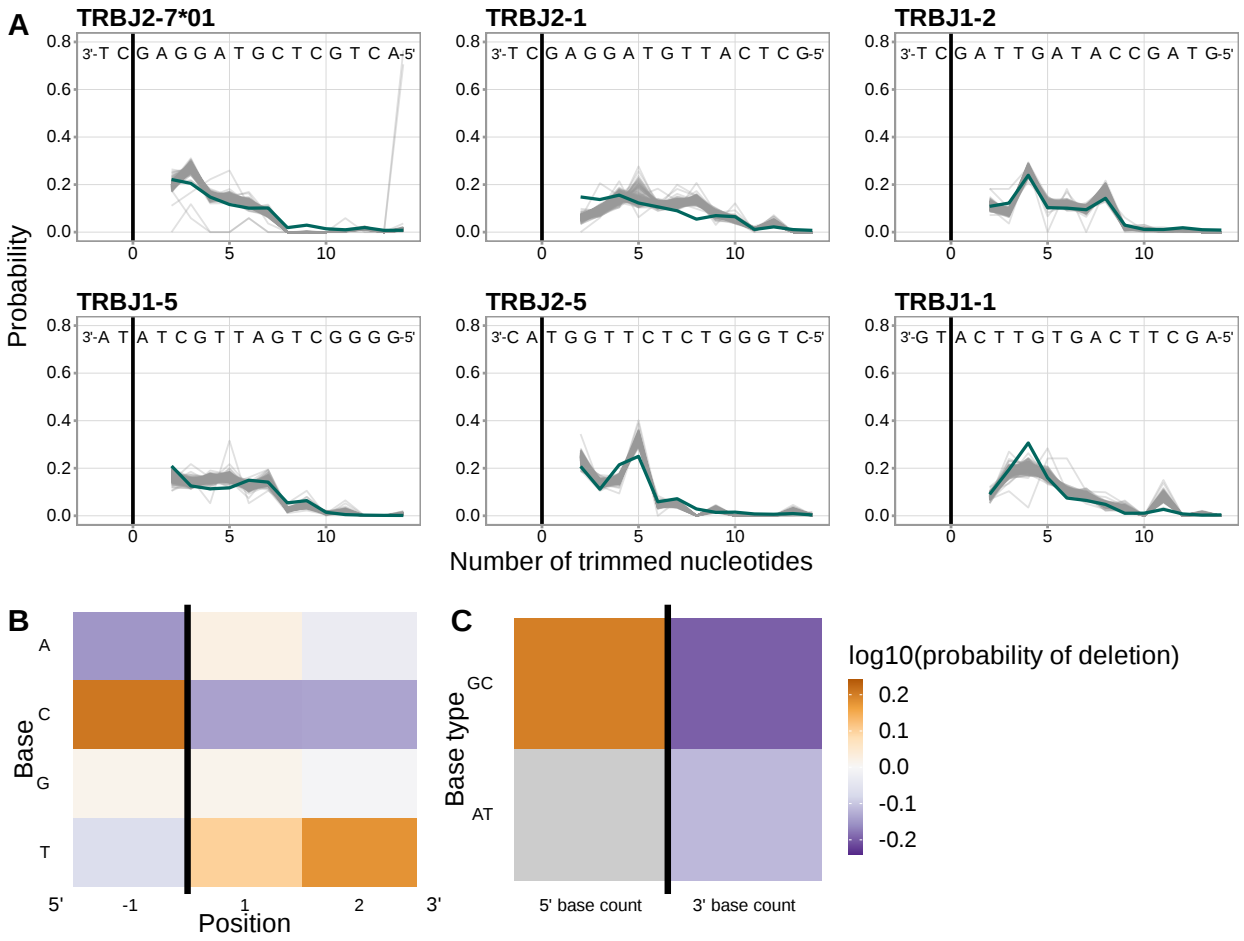

Figure C.7: (**A**) Inferred trimming profiles (shown in blue) using the *1x2 motif + two-side base-count beyond* model trained using J-gene sequences have good predictive accuracy for the most frequently used J-genes. Gene-specific trimming profiles for each individual in the training data set are shown in gray. (**B**) Position-weight-matrix of the local sequence context dependence of J-gene trimming probabilities consisting of 1 nucleotides 5' of the trimming site and 2 nucleotides 3' of the trimming site from fitting the *1x2 motif + two-side base-count beyond* model using J-gene sequences. Positions 5' and 3' of the trimming site have a strong effect on the probability of trimming. (**C**) Inferred *two-side base-count beyond* model coefficients from fitting the *1x2 motif + two-side base-count beyond* model using J-gene sequences suggest that the count of GC bases 5' of the motif has a strong positive effect on the trimming probability whereas the count of GC and/or AT bases 3' of the motif has a negative effect. The count of AT nucleotides 5' of the motif (shown in gray) was not included in this model. The black vertical line corresponds to the trimming site. Each inferred coefficient is given as the change in log10 odds of trimming at a given site resulting from an increase in the feature value, given that all other features are held constant.

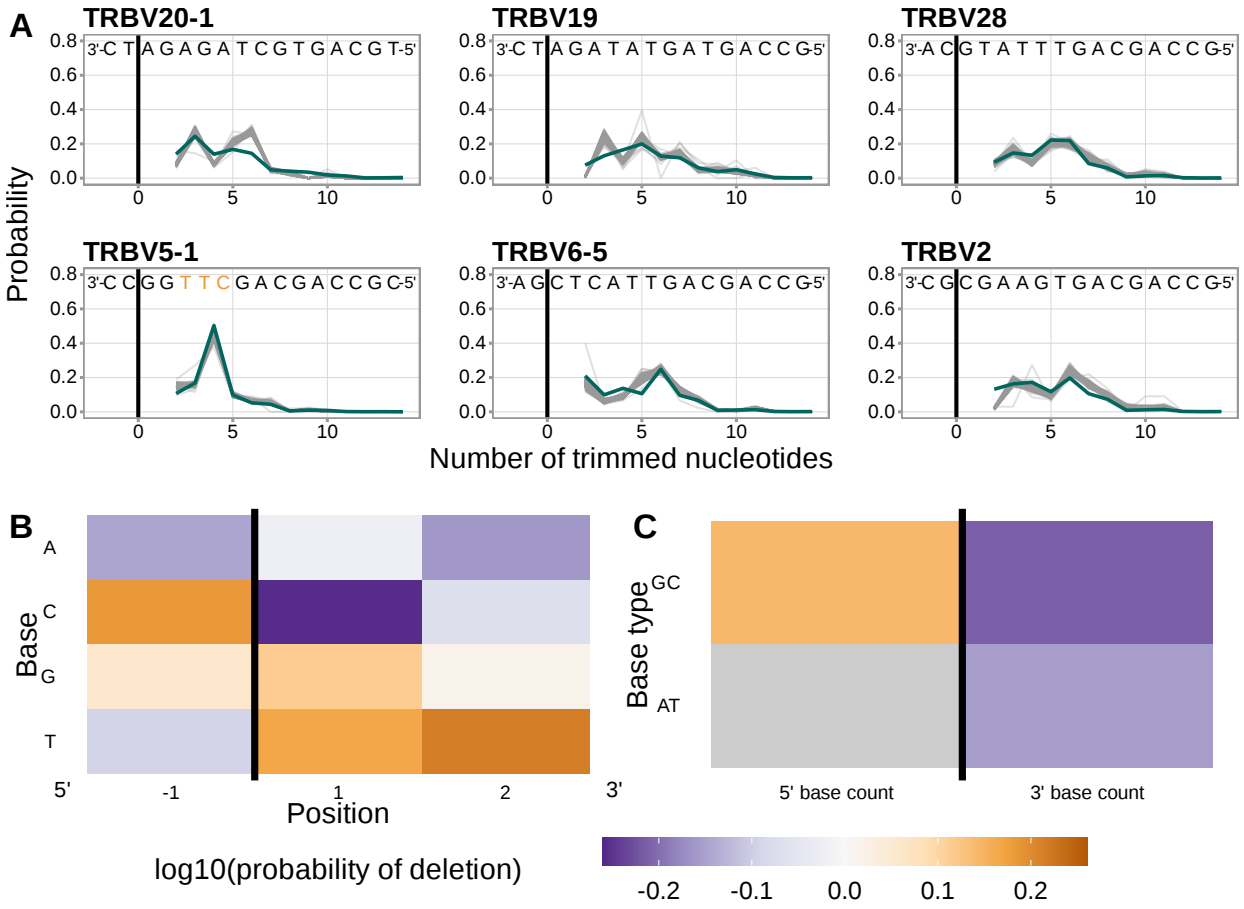

Figure C.8: (**A**) Inferred trimming profiles (shown in blue) using the *1x2 motif + two-side base-count beyond* model trained using productive V-gene sequences have good predictive accuracy for the most frequently used V-genes. Gene-specific trimming profiles for each individual in the training data set are shown in gray. (**B**) Position-weight-matrix of the local sequence context dependence of V-gene trimming probabilities consisting of 1 nucleotides 5' of the trimming site and 2 nucleotides 3' of the trimming site from fitting the *1x2 motif + two-side base-count beyond* model using productive V-gene sequences. Positions 5' and 3' of the trimming site have a strong effect on the probability of trimming. (**C**) Inferred *two-side base-count beyond* model coefficients from fitting the *1x2 motif + two-side base-count beyond* model using productive V-gene sequences suggest that the count of GC bases 5' of the motif has a strong positive effect on the trimming probability whereas the count of GC and/or AT bases 3' of the motif has a negative effect. The count of AT nucleotides 5' of the motif (shown in gray) was not included in this model. The black vertical line corresponds to the trimming site. Each inferred coefficient is given as the change in log10 odds of trimming at a given site resulting from an increase in the feature value, given that all other features are held constant.

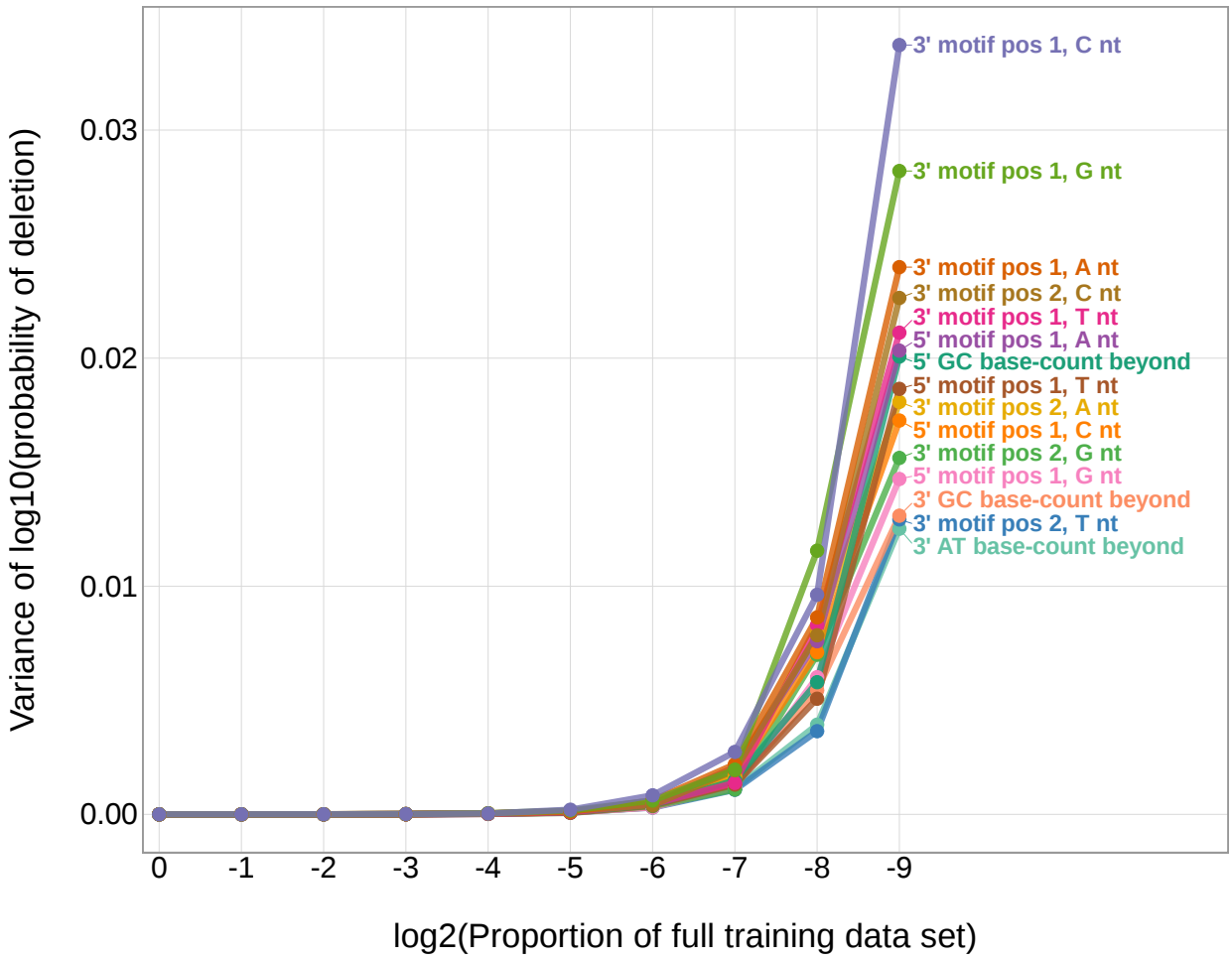

Figure C.9: The magnitudes of the inferred coefficients from the *1x2 motif + two-side base-count beyond* model have minimal variance when changing the number of sequences included in the training data set. The original V-gene training data set contains 21,193,153 sequences. When sub-sampling the original V-gene training data set and re-training the model, the inferred coefficients are stable until the size of the training data set reaches around 82,800 sequences (e.g. log2(Proportion of the full training data set) is equal to -8) or below.

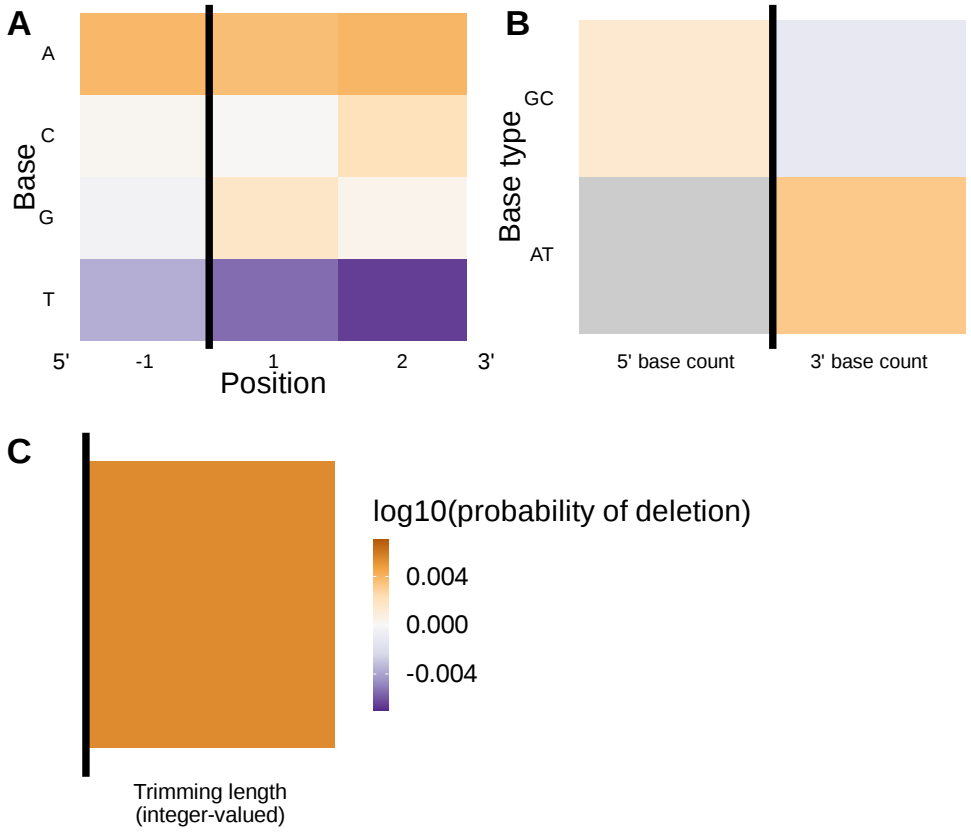

Figure C.10: Inferred SNP-interaction coefficients from fitting the *1x2 motif + two-side base-count beyond proportion + length* SNP interaction model. (**A**) Inferred interaction coefficients between rs41298872 SNP genotype and *motif* parameters for 1 nucleotide position 5' of the trimming site and 2 nucleotide positions 3' of the trimming site. The interaction coefficients between the SNP genotype and the presence of T nucleotides (at all positions in the motif) are significant. (**B**) Inferred interaction coefficients between rs41298872 SNP genotype and *two-side base-count beyond* model coefficients. The interaction coefficient between the SNP genotype and the count of AT nucleotides 5' of the motif (shown in gray) was not included in this model. None of the interaction coefficients are significant. (**C**) Inferred interaction coefficients between rs41298872 SNP genotype and the *length* coefficient The black vertical line corresponds to the trimming site. Each inferred coefficient is given as the change in log10 odds of trimming at a given site resulting from an increase in the feature value, given that all other features are held constant.

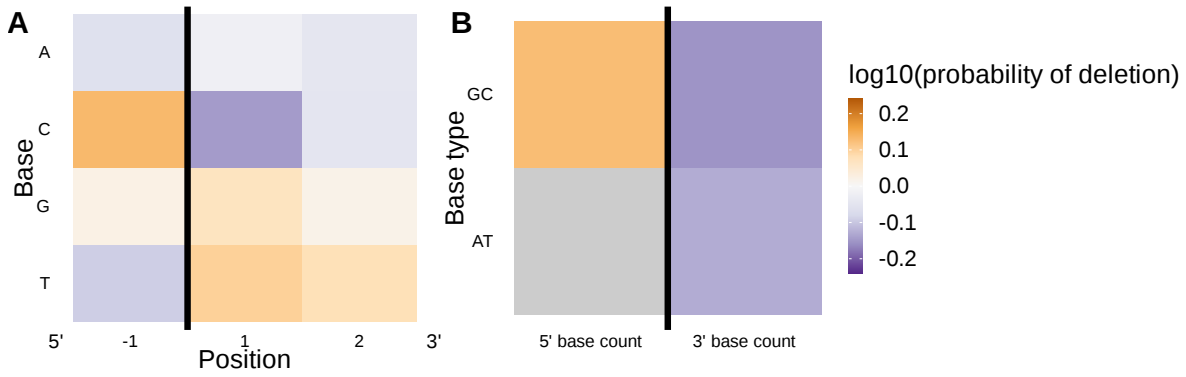

Figure C.11: Differing methods of sequence annotation have little to no effect on the model fit or performance. (**A**) Position-weight-matrix of the local sequence context dependence of V-gene trimming probabilities consisting of 1 nucleotides 5' of the trimming site and 2 nucleotides 3' of the trimming site from fitting the *1x2 motif + two-side base-count beyond* model using parsimony-annotated sequences. (**B**) Inferred *two-side base-count beyond* model coefficients from fitting the *1x2 motif + two-side base-count beyond* model using parsimony-annotated sequences. These inferred coefficients are highly similar to the original *1x2 motif + two-side base-count beyond* model trained using IGoR-annotated sequences (Figure 3.4). The count of AT nucleotides 5' of the motif (shown in gray) was not included in this model. The black vertical line corresponds to the trimming site. Each inferred coefficient is given as the change in log10 odds of trimming at a given site resulting from an increase in the feature value, given that all other features are held constant.

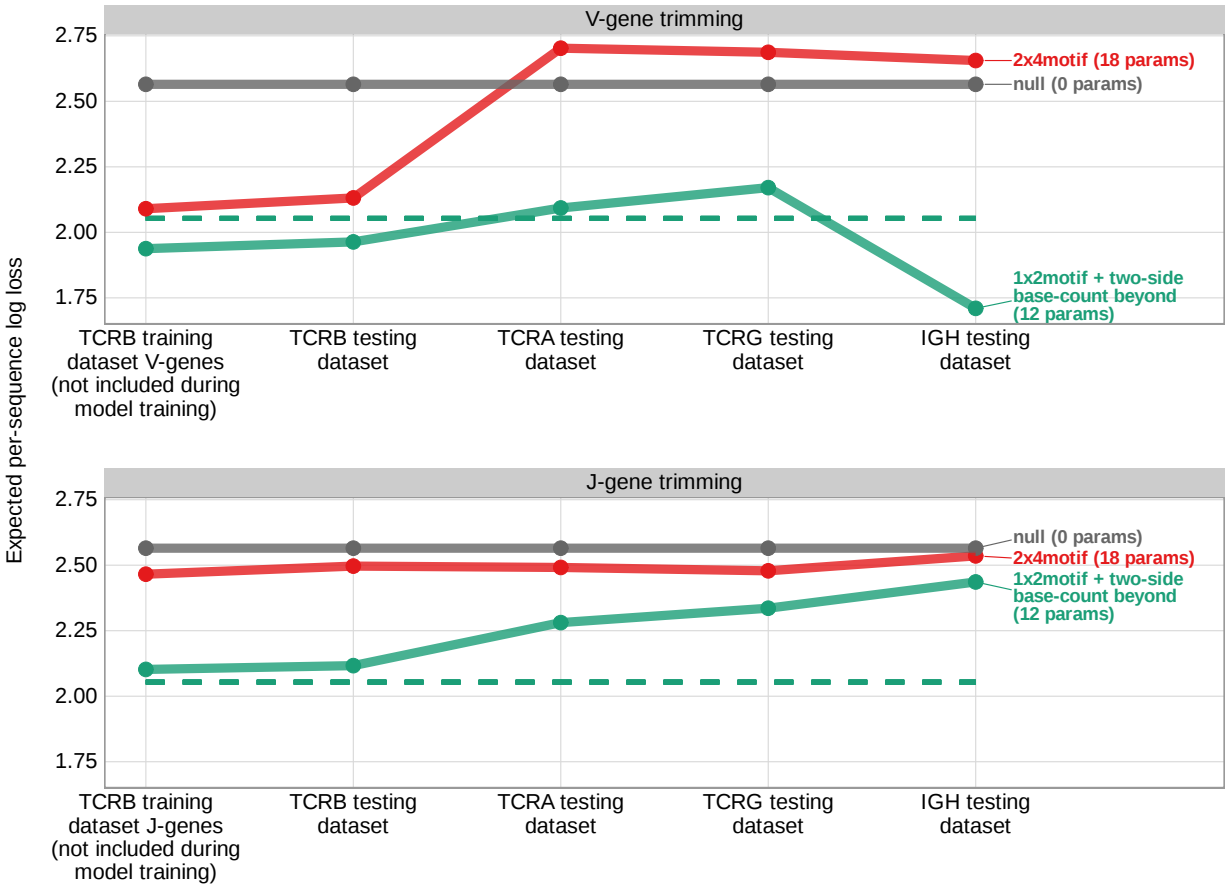

Figure C.12: Model performance is similar for productive sequences compared to non-productive sequences from each testing data set. Expected per-sequence conditional log loss computed for various models using the TCRβ V-gene training data set and productive V- and J-gene sequences from several independent testing data sets. Each model was trained using the full non-productive TCRβ V-gene training data set. Lower expected per-sequence log loss corresponds to a better model fit. The *1x2 motif + two-side base-count beyond* model has the best model fit and generalizability across all testing data sets. The horizontal dashed line corresponds to the expected per-sequence log loss of the *1x2 motif + two-side base-count beyond* model computed for V-gene trimming using the non-productive TCRβ V-gene training data set.

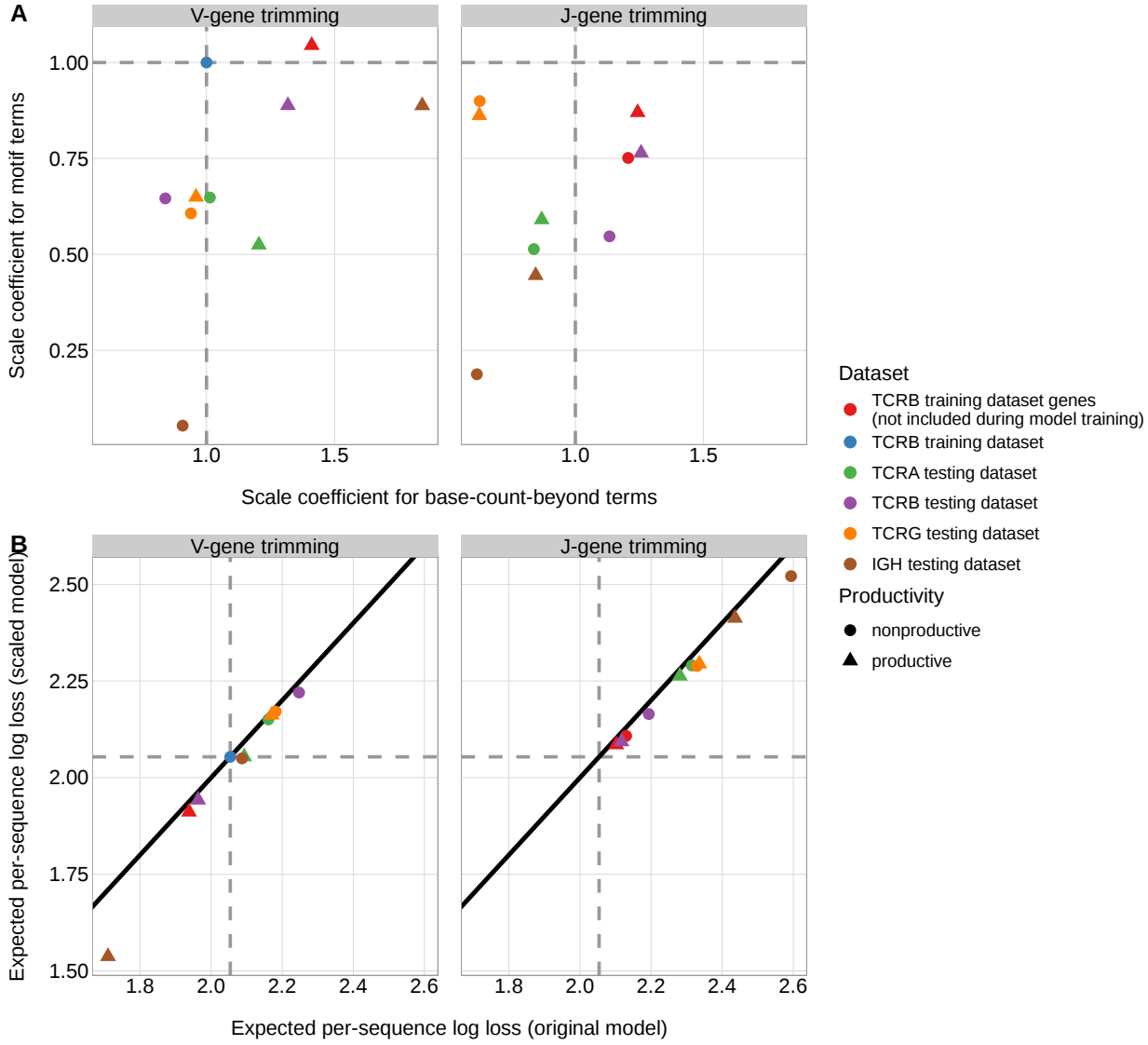

Figure C.13: The weight of the *1x2 motif* and *two-side base-count beyond* model terms varies by data set. (**A**) The scale coefficient for the *two-side base-count beyond* model terms is larger than the *motif* scale coefficient for every data set. (**B**) The expected per-sequence conditional log loss of each of these new models is only slightly better compared to the original *1x2 motif + two-side base-count beyond* model. Horizontal and vertical dashed gray lines correspond to the TCRβ V-gene training data set. The black solid line corresponds to the $y = x$ line.

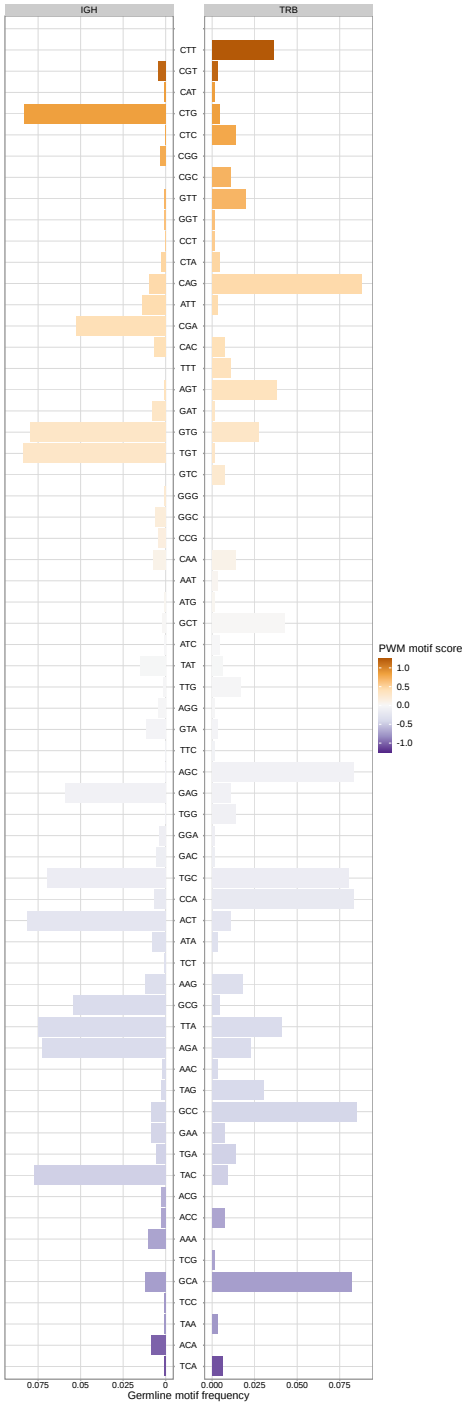

Figure C.14: Sequence motifs appear at varying frequencies within the germline *TRB* and *IGH* genes. Sequence motifs that lead to a large increase in trimming probabilities in the model (e.g. a large, positive PWM motif score) appear at relatively low frequencies within the germline IGH genes. This may explain the weakness of the motif-related signal within *IGH* data sets (Figure C.13).

Table C.1: Summary of all notation used in our trimming modeling

Variable	Description
General notation	
I	set of all individuals
i	index for an individual in the set I of all individuals
K_i	total number of TCRs in the repertoire of individual i
k	index of a sequence in the TCR repertoire of individual i
S	random variable that represents the gene sequence
σ	general notation for a gene allele group sequence oriented 5'-to-3'
σ_V	V-gene allele group sequence ("top" strand oriented 5'-to-3')
σ_J	J-gene allele group sequence ("bottom" strand oriented 5'-to-3')
N	random variable that represents the number of deleted nucleotides
n	number of deleted nucleotides from the 3'-side of a gene sequence
L	lower bound of "reasonable" trimming amounts, we have defined $L = 2$
U	upper bound of "reasonable" trimming amounts, we have defined $U = 14$
$C^{(i)}(\sigma)$	the number of TCRs that use gene allele group σ in the sampled repertoire of individual i
$C^{(i)}(n, \sigma)$	the number of TCRs that have gene allele group σ and n nucleotides deleted in the sampled repertoire of individual i
N'	set of all "reasonable" trimming amounts; $N' = \{2, \dots 14\}$
$P_{\text{emp}}(N = n \mid S = \sigma, i)$	empirical conditional probability density function (3.1)
Motif parameter-specific notation	
a	non-negative integer value that represents the number of nucleotides 5' of the trimming site to be included in the "trimming motif"
b	non-negative integer value that represents the number of nucleotides 3' of the trimming site to be included in the "trimming motif"
$\{\sigma(n+j)\}_{j=-a}^{b-1}$	"trimming motif" sequence (D.1)
$\beta_{js}^{\texttt{motif}}$	(log) position weight matrix coefficient for trimming motif position $j \in \{-a, \dots, b-1\}$ and nucleotide $s \in \{A, T, C, G\}$
$\boldsymbol{\beta}^{\texttt{motif}}$	set of all *motif* coefficients $\beta_{js}^{\texttt{motif}}$ for all positions $j \in \{-a, \dots, b-1\}$ and nucleotide $s \in \{A, T, C, G\}$
$f(n, \sigma; \boldsymbol{\beta}^{\texttt{motif}}, a, b)$	*motif*-specific covariate function (D.2)
Base-count-beyond parameter-specific notation	

c	non-negative integer value that represents the number of nucleotides 5' of the trimming site to be included in the 5' base-count-beyond the "trimming motif"
$C^{\mathtt{AT}}(x)$	count of nucleotides that are A or T in an arbitrary sequence x
$C^{\mathtt{GC}}(x)$	count of nucleotides that are G or C in an arbitrary sequence x
$\mathsf{seq}_5(n, \sigma, a, c)$	the nucleotide sequence 5' of the trimming site, beyond the "trimming motif" (D.3)
$\mathsf{seq}_3(n, \sigma, b)$	the nucleotide sequence 3' of the trimming site, beyond the "trimming motif" (D.4)
$\beta_5^{\mathtt{AT}}$ and $\beta_3^{\mathtt{AT}}$	*base-count-beyond* model coefficients for the 5' and 3' sequence base-counts of A and T nucleotides beyond the trimming motif
$\boldsymbol{\beta}^{\mathtt{AT}}$	set of AT-*base-count-beyond* model coefficients (includes $\beta_5^{\mathtt{AT}}$ and $\beta_3^{\mathtt{AT}}$)
$\beta_5^{\mathtt{GC}}$ and $\beta_3^{\mathtt{GC}}$	*base-count-beyond* model coefficients for the 5' and 3' sequence base-counts of G and C nucleotides beyond the trimming motif
$\boldsymbol{\beta}^{\mathtt{GC}}$	set of GC-*base-count-beyond* model coefficients (includes $\beta_5^{\mathtt{GC}}$ and $\beta_3^{\mathtt{GC}}$)
$f(n, \sigma; \boldsymbol{\beta}^{\mathtt{AT}}, \boldsymbol{\beta}^{\mathtt{GC}}, a, b, c)$	*base-count-beyond*-specific covariate function (D.2)
DNA-shape parameter-specific notation	
$\mathsf{seq}_{\mathrm{expd}}(n, \sigma, a, b)$	"expanded trimming sequence window" (D.6); consists of the "trimming motif" sequence extended by two nucleotides in both the 5' and 3' direction
$\mathtt{E}$	nucleotide electrostatic potential
$\mathtt{W}$	nucleotide minor groove width
$\mathtt{P}$	nucleotide propeller twist
$\mathtt{R}$	di-nucleotide roll
$\mathtt{H}$	di-nucleotide helical twist
$\mathsf{shape}^{\mathtt{u}}(j, \mathsf{seq}_{\mathrm{expd}}(n, \sigma, a, b))$	measure of nucleotide shape $u \in \{\mathtt{E}, \mathtt{W}, \mathtt{P}\}$ for the nucleotide at position $j \in \{-a, \ldots, b-1\}$ within the "expanded trimming sequence window" $\mathsf{seq}_{\mathrm{expd}}(n, \sigma, a, b)$
$\mathsf{shape}^{\mathtt{v}}(d, \mathsf{seq}_{\mathrm{expd}}(n, \sigma, a, b))$	measure of di-nucleotide shape $v \in \{\mathtt{R}, \mathtt{H}\}$ for the di-nucleotide at position $d \in \{-a+1, \ldots, b-1\}$ within the "expanded trimming sequence window" $\mathsf{seq}_{\mathrm{expd}}(n, \sigma, a, b)$
$\beta_{uj}^{\mathtt{shape}}$	DNA-shape coefficients for nucleotide shape type $u \in \{\mathtt{E}, \mathtt{W}, \mathtt{P}\}$ and "expanded trimming sequence window" nucleotide position $j \in \{-a, \ldots, b-1\}$
$\beta_{vd}^{\mathtt{shape}}$	DNA-shape coefficients for di-nucleotide shape type $v \in \{\mathtt{R}, \mathtt{H}\}$ and "expanded trimming sequence window" di-nucleotide position $d \in \{-a+1, \ldots, b-1\}$
$\boldsymbol{\beta}^{\mathtt{shape}}$	set of all nucleotide and di-nucleotide DNA-shape coefficients

$f(n, \sigma; \boldsymbol{\beta}^{\texttt{shape}}, a, b)$	DNA-shape-specific covariate function (D.7)
Length parameter-specific notation	
$\beta^{\texttt{ldist}}$	*length* specific model coefficient
$f(n, \sigma; \beta^{\texttt{ldist}})$	*length*-specific covariate function
Modeling notation	
$f(n, \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$	example model covariate function including *motif* and *base-count-beyond* model parameters (3.2)
$P(n \mid \sigma; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$	conditional logit model formulation using the *motif* and *base-count-beyond* model covariate function (3.3)
$\log L(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$	aggregated log likelihood for the conditional logit model; this likelihood function is un-weighted (3.4) and gives every observation uniform treatment in the likelihood
$P_{\text{sample}}(N = n, S = \sigma)$	sampling procedure for the construction of the expected likelihood
$\log L_{\text{expected}}(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$	expected log likelihood for the conditional logit model; this likelihood function (3.5) weights each observation by its sampling probability, $P_{\text{sample}}(N = n, S = \sigma)$
$\log L_{\text{emp}}(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$	expected log likelihood for the conditional logit model; this likelihood function (3.7) weights each observation by its sampling probability from the empirical joint PDF (3.6)
$P_{\text{marg}}(\sigma)$	empirical average per-gene-allele-group frequency used in formulating a subject-independent gene sampling procedure (3.8)
$\log L_W(\boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$	expected log likelihood for the conditional logit model; this likelihood function (3.9) weights each observation using a subject-independent gene sampling procedure (3.8)
Model evaluation notation	
$\mathcal{M}$	an arbitrary model trained on a specified training data set
$\mathbf{V}$	full V-gene data set
$\mathbf{J}$	full J-gene data set
$\mathbf{H}$	arbitrary held-out data set
$P(\mathbf{H})$	probability of the arbitrary held-out data set (D.9)
$\ell(\mathcal{M} \mid \mathbf{H})$	expected per-sequence conditional log loss (3.11) of a trained model $\mathcal{M}$ evaluated on a data set $\mathbf{H}$
$E[\ell(\mathcal{M})]$	expected per-sequence conditional log loss across 20 random held-out data sets (D.10)
$\text{RMSE}(\sigma, \mathcal{M}, \mathbf{V})$	per-gene mean squared error (D.11) for a gene σ using a model $\mathcal{M}$ trained using the V-gene training data set $\mathbf{V}$
Coefficient evaluation notation	
$T(\hat{\beta})$	test statistic (3.12) for evaluating the significance of a single inferred coefficient $\hat{\beta}$
X	set of single-nucleotide polymorphisms (SNPs) within the gene encoding the Artemis protein that were previously identified to be associated with increasing the extent of trimming [2]

g_{ix}	number of minor alleles in the genotype of an individual $i \in I$ for SNP $x \in X$
$\{\boldsymbol{\beta}_x^{\texttt{motif}}, \boldsymbol{\beta}_x^{\texttt{AT}}, \boldsymbol{\beta}_x^{\texttt{GC}}\}$	set of interaction coefficients between each model parameter and the SNP x genotype

Appendix D

SUPPLEMENTARY METHODS FOR CHAPTER 3

D.1 Extended parameter description

Defining the "trimming motif" and position-weight-matrix weight for a given gene and trimming site

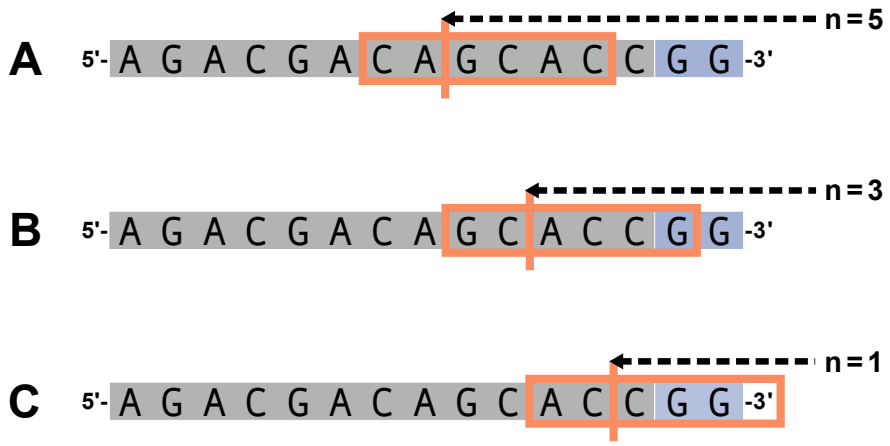

Figure D.1: Let $a = 2$ and $b = 4$. The 6 nucleotide trimming motif given by $(\mathrm{D}.1)$ is shown in the orange box and the trimming site is shown by the vertical orange line. An arbitrary gene sequence is highlighted in gray and the two possible P-nucleotides are highlighted in purple. (**A**) For $n = 5$, the 6 nucleotide trimming motif will not contain P-nucleotides. (**B**) For $n = 3$, the 6 nucleotide trimming motif will contain one P-nucleotide. (**C**) For $n = 1$, the trimming motif will contain two P-nucleotides and will be "incomplete" (contain less than 6 nucleotides).

Existing probabilistic models of nucleotide trimming using repertoire sequencing data have shown that the local nucleotide context around the trimming site, which we refer to as the "trimming motif", do a surprisingly good job of predicting the distribution of trimming

lengths for a variety of genes [15]. This simple position-weight-matrix style model uses a trimming motif containing two nucleotides 5' of the trimming site and four nucleotides 3' of the trimming site to predict the probability of trimming at that site. In practice, we can define the trimming motif to be any size. Let a and b be non-negative integer values that represent the number of nucleotides 5' and 3' of the trimming site, respectively. Together, these $a + b$ nucleotides will compose the trimming motif. For a gene allele group sequence σ and a number of deleted nucleotides n, let $\sigma(n + j)$ represent the nucleotide identity at the trimming motif position $j \in \{-a, \ldots, b-1\}$ where positions $j < 0$ represent motif positions 5' of the trimming site and positions $j \geq 0$ represent motif positions 3' of the trimming site. As such, the trimming motif sequence is given by

$$\{\sigma(n + j)\}_{j=-a}^{b-1}. \tag{D.1}$$

Depending on n, this trimming motif may or may not include P-nucleotides. For example, for $n \geq b$, the b 3' trimming motif nucleotides will include the b deleted gene sequence nucleotides 3' of the trimming site (and no P-nucleotides) (Figure D.1A). Since we are assuming that the initial hairpin nick occurs at the $+2$ position, there will be two P-nucleotides present in the 5'-to-3' gene sequence. For $b - 2 \leq n < b$, where the 2 represents the total P-nucleotide count in the full sequence, P-nucleotides will be included in the trimming motif sequence. Specifically, the b total 3' trimming motif nucleotides will include $b - n$ P-nucleotides and n deleted gene sequence nucleotides (Figure D.1B-C). Likewise, as a result of the $+2$ hairpin nick position assumption, TCRs that have $n < b - 2$ will not have a full, $(a + b)$-length nucleotide trimming motif (Figure D.1C). For these "off-the-end" motif cases, we assign zero influence to the missing nucleotides during model fitting.

With this trimming motif, let $\beta_{js}^{\texttt{motif}}$ be a (log) position-weight-matrix coefficient for trimming motif position $j \in \{-a, \ldots, b-1\}$ and nucleotide $s \in \{A, T, C, G\}$. We can define

an un-normalized position-weight-matrix weight

$$f(n, \sigma; \boldsymbol{\beta}^{\texttt{motif}}, a, b) := \sum_{j=-a}^{b-1} \beta^{\texttt{motif}}_{j\sigma(n+j)} \tag{D.2}$$

that will serve as a *motif*-specific model covariate function in subsequent modeling. As described above, since we are considering "off-the-end" motif cases, $\sigma(n+j)$ represent the nucleotide identity at sequence position j where positions $j < 0$ represent sequence positions 5' of the trimming site and positions $j \geq 0$ represent sequence positions 3' of the trimming site.

AT and GC base-count-beyond the trimming motif

For an arbitrary sequence x, we can count the number of AT and GC nucleotides within the sequence as

$$C^{\texttt{AT}}(x) = C^{\texttt{A}}(x) + C^{\texttt{T}}(x)$$

and

$$C^{\texttt{GC}}(x) = C^{\texttt{G}}(x) + C^{\texttt{C}}(x),$$

respectively.

Because the count of AT or GC nucleotides within the sequences 5' and 3' of the trimming site may influence the probability of trimming differently, we will calculate the counts separately. We will not include nucleotides that were already included in the *motif* parameterization. As above, for a gene allele group sequence σ and a number of deleted nucleotides n, let $\sigma(n+j)$ represent the nucleotide identity at sequence position j where positions $j < 0$ represent sequence positions 5' of the trimming site and positions $j \geq 0$ represent sequence positions 3' of the trimming site. Let c be a non-negative integer value that represents the number of nucleotides 5' of the trimming site that will be included in the 5' nucleotide counts

(Figure D.2). Recall that a is a non-negative integer value that represents the number of nucleotides 5' of the trimming site that are included in the "trimming motif" described in the previous section. As such, the nucleotide sequence 5' of the trimming site, beyond the "trimming motif", is given by

$$\mathsf{seq}_5(n, \sigma, a, c) = \{\sigma(n + j)\}_{j=(a+1)}^{(a+c)}. \tag{D.3}$$

Within this sequence $\mathsf{seq}_5(n, \sigma, a, c)$, we can count the number of AT and GC nucleotides as

$$C^{\mathtt{AT}}(\mathsf{seq}_5(n, \sigma, a, c)) = C^{\mathtt{A}}(\mathsf{seq}_5(n, \sigma, a, c)) + C^{\mathtt{T}}(\mathsf{seq}_5(n, \sigma, a, c))$$

and

$$C^{\mathtt{GC}}(\mathsf{seq}_5(n, \sigma, a, c)) = C^{\mathtt{G}}(\mathsf{seq}_5(n, \sigma, a, c)) + C^{\mathtt{C}}(\mathsf{seq}_5(n, \sigma, a, c)),$$

respectively.

To count the number of AT and GC nucleotides in the sequence 3' of the trimming site, we will include all nucleotides located 3' of the trimming site that are beyond the "trimming motif." However, because we are interested in using GC nucleotide content in both directions of the wider sequence as a proxy for the capacity for sequence-breathing and since sequence-breathing is only relevant for nucleotides that are paired, we will not include the nucleotides within the 3' single-stranded-overhang when counting 3' AT and GC nucleotides (Figure D.2). Since we are assuming that the initial hairpin nick occurs at the $+2$ position leading to a 4 nucleotide long 3' single-stranded-overhang, for $n > 2$, the nucleotide sequence 3' of the trimming site, beyond the "trimming motif", is given by

$$\mathsf{seq}_3(n, \sigma, b) = \begin{cases} \{\sigma(n + j)\}_{j=3-n}^{-b} & \text{if } (n-3) \geq b \\ \{\} & \text{if } (n-3) < b \end{cases} \tag{D.4}$$

where b is a non-negative integer value that represents the number of nucleotides 3' of the trimming site that are included in the "trimming motif" described in the previous section. For $(n-3) < b$, all nucleotides 3' of the trimming site are considered single-stranded and, thus, no nucleotides will be included in the sequence used to calculate the AT and GC base-counts (Figure D.2C). Within this sequence $\mathsf{seq}_3(n, \sigma, b)$, we can count the number of AT and GC nucleotides as

$$C^{\mathtt{AT}}(\mathsf{seq}_3(n, \sigma, b)) = C^{\mathtt{A}}(\mathsf{seq}_3(n, \sigma, b)) + C^{\mathtt{T}}(\mathsf{seq}_3(n, \sigma, b))$$

and

$$C^{\mathtt{GC}}(\mathsf{seq}_3(n, \sigma, b)) = C^{\mathtt{G}}(\mathsf{seq}_3(n, \sigma, b)) + C^{\mathtt{C}}(\mathsf{seq}_3(n, \sigma, b)),$$

respectively. As defined, these GC and AT base-counts for the 3' sequence are dependent on sequence length and provide a parameterization of both GC nucleotide content in both directions of the wider sequence and length.

With these 5' and 3' base counts, we can define $\beta_5^{\mathtt{AT}}$, $\beta_3^{\mathtt{AT}}$, $\beta_5^{\mathtt{GC}}$, and $\beta_3^{\mathtt{GC}}$ to be *base-count-beyond* model coefficients for 5' and 3' sequence base-counts of AT and GC beyond the "trimming motif", respectively. With these coefficients, we can define a *base-count-beyond* covariate function for each trimming site n and gene σ:

$$\begin{aligned} f(n, \sigma; \boldsymbol{\beta}^{\mathtt{AT}}, \boldsymbol{\beta}^{\mathtt{GC}}, a, b, c) \coloneqq \beta_5^{\mathtt{AT}} \cdot C^{\mathtt{AT}}(\mathsf{seq}_5(n, \sigma, a, c)) + \beta_3^{\mathtt{AT}} \cdot C^{\mathtt{AT}}(\mathsf{seq}_3(n, \sigma, b)) \\ + \beta_5^{\mathtt{GC}} \cdot C^{\mathtt{GC}}(\mathsf{seq}_5(n, \sigma, a, c)) + \beta_3^{\mathtt{GC}} \cdot C^{\mathtt{GC}}(\mathsf{seq}_3(n, \sigma, b)). \end{aligned} \tag{D.5}$$

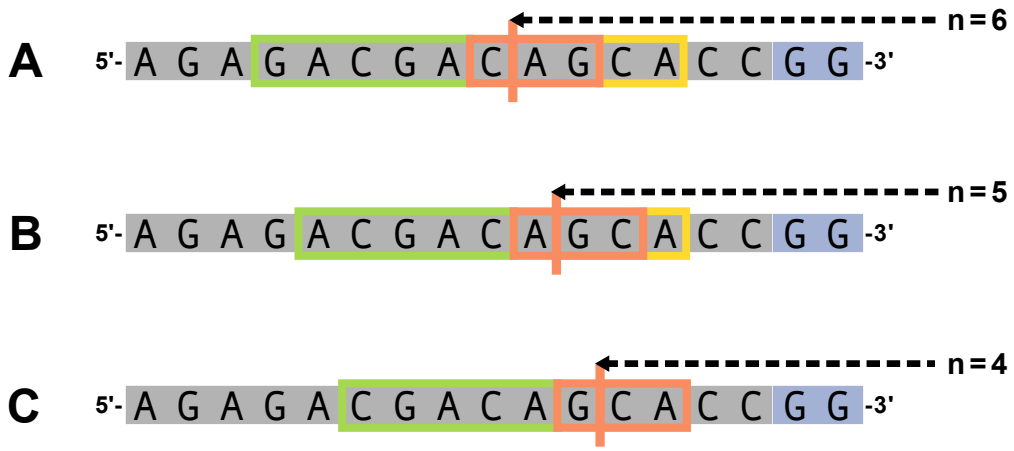

Figure D.2: Let $a = 1$, $b = 2$, and $c = 5$. An arbitrary gene sequence is highlighted in gray and the two possible P-nucleotides are highlighted in purple. The trimming site is shown by the vertical orange line and the "trimming motif", as defined in $(\mathrm{D}.1)$, is shown by the orange box. The c nucleotides included in the count of AT and GC nucleotides 5' of the trimming site, beyond the "trimming motif", are expressed by $(\mathrm{D}.3)$ and are shown in the green box. The nucleotides included in the count of AT and GC nucleotides 3' of the trimming site, beyond the "trimming motif", are expressed by $(\mathrm{D}.4)$ and are shown in the yellow box. As described in the text, we are assuming that the initial hairpin nick occurs at the +2 position leading to a 4 nucleotide long 3' single-stranded-overhang. We exclude these single-stranded nucleotides in the 3' base-count-beyond sequence. In this figure, the 4 nucleotides nearest to the 3' side of each sequence (this includes the two P-nucleotides and the two 3'-most gene sequence nucleotides) are considered single-stranded and will not be included in the 3' base-count-beyond sequence. (**A**) For $n = 6$, 2 nucleotides 3' of the trimming site will be used in the 3' sequence base-counts. (**B**) For $n = 5$, 1 nucleotide 3' of the trimming site will be used in the 3' sequence base-counts. (**C**) For $n = 4$, all nucleotides 3' of the trimming site are considered single-stranded and, thus, no nucleotides will be used for the 3' sequence base-counts.

DNA-shape around the trimming site

Methods have been previously developed to estimate DNA-shape features at a single-nucleotide position using the sequence context of two neighboring nucleotides on both sides of the nucleotide of interest [98, 99]. As such, these methods use a sliding-pentamer model, centered at each nucleotide of interest, to derive the structural features of nucleotides within a sequence window of any length. These structural features include estimations of electrostatic potential (E), minor groove width (W), and propeller twist (P) for each nucleotide in the

sequence window and estimations of roll (R) and helical twist (H) for each di-nucleotide pair in the sequence window. For simplicity, we will use the term "DNA-shape parameters" to refer to all five of these structural features.

For our purposes, we can define a "trimming sequence window" of size $a+b$, as introduced in the "trimming motif" section with (D.1), where a and b are non-negative integer values that represent the number of nucleotides 5' and 3' of the trimming site, respectively. In order to estimate the DNA-shape for all nucleotides within this window, we will expand the "trimming sequence window" by two nucleotides on both sides such that there are $a+2$ nucleotides 5' and $b+2$ nucleotides 3' of the trimming site included in an "expanded trimming sequence window." For a gene allele group sequence σ and a number of deleted nucleotides n, let $\sigma(n+j)$ represent the nucleotide identity at the "expanded trimming sequence window" position $j \in \{-(a+2), \ldots, (b+2)-1\}$ where positions $j < 0$ represent expanded trimming sequence window positions 5' of the trimming site and positions $j \geq 0$ represent expanded trimming sequence window positions 3' of the trimming site. As such, the expanded trimming sequence window is given by

$$\mathsf{seq}_{\text{expd}}(n, \sigma, a, b) := \{\sigma(n+j)\}_{j=-(a+2)}^{(b+2)-1}. \tag{D.6}$$

Depending on n, this expanded trimming sequence window may or may not include P-nucleotides. For example, for $n \geq (b+2)$, the $(b+2)$ 3' expanded trimming sequence window nucleotides will include the $(b+2)$ deleted gene sequence nucleotides 3' of the trimming site (and no P-nucleotides) (Figure D.3A). For $b \leq n < b+2$, the $(b+2)$ 3' expanded trimming sequence window nucleotides will include $(b+2)-n$ P-nucleotides and n deleted gene sequence nucleotides (Figure D.3B). Since we are assuming that the initial hairpin nick occurs at the $+2$ position, TCRs that have $n < b$ will not have a full, $(a+b+4)$-length nucleotide expanded trimming sequence window (Figure D.3C). The sliding-pentamer model [98, 99] requires a full

pentamer for estimating the DNA-shape of each base of interest, and, thus, for these "off-the-end" expanded trimming sequence window cases, we cannot estimate DNA-shape parameters for all nucleotides within the trimming sequence window. As such, when estimating DNA-shape parameters, we must choose b such that $b \leq n$ for all trimming lengths n in the data set.

For each nucleotide position $j \in \{-a, ..., b-1\}$ within the expanded trimming sequence window $\mathsf{seq}_{\text{expd}}(n, \sigma, a, b)$, we can estimate the nucleotide electrostatic potential, $\mathsf{shape}^{\mathtt{E}}(j, \mathsf{seq}_{\text{expd}}(n, \sigma, a, b))$, minor groove width, $\mathsf{shape}^{\mathtt{W}}(j, \mathsf{seq}_{\text{expd}}(n, \sigma, a, b))$, and propeller twist, $\mathsf{shape}^{\mathtt{P}}(j, \mathsf{seq}_{\text{expd}}(n, \sigma, a, b))$. We then standardize the estimated values for each shape type. We can define $\beta_{uj}^{\mathtt{shape}}$ to be a nucleotide shape model coefficient for nucleotide shape type $u \in \{\mathtt{E}, \mathtt{W}, \mathtt{P}\}$ and trimming sequence window nucleotide position $j \in \{-a, ..., b-1\}$. Let $d \in \{-a+1, ..., b-1\}$ be the location of each di-nucleotide in the trimming sequence window such that $d = 0$ represents the location of the trimming site, $d < 0$ represents di-nucleotide positions 5' of the trimming site, and $d > 0$ represents di-nucleotide positions 3' of the trimming site. For each di-nucleotide $d \in \{-a+1, ..., b-1\}$ within the expanded trimming sequence window $\mathsf{seq}_{\text{expd}}(n, \sigma, a, b)$, we can estimate the di-nucleotide roll, $\mathsf{shape}^{\mathtt{R}}(d, \mathsf{seq}_{\text{expd}}(n, \sigma, a, b))$ and helical twist, $\mathsf{shape}^{\mathtt{H}}(d, \mathsf{seq}_{\text{expd}}(n, \sigma, a, b))$. As above, we then standardize the estimated values for each di-nucleotide shape type. We can define $\beta_{vd}^{\mathtt{shape}}$ to be a di-nucleotide shape model coefficient for di-nucleotide shape type $v \in \{\mathtt{R}, \mathtt{H}\}$ and trimming sequence window di-nucleotide position $d \in \{-a+1, ..., b-1\}$. We use the R package `DNAshapeR` [99] to estimate these DNA-shape parameters for each trimming sequence window. With these standardized DNA-shape estimates, we can define a DNA-shape

covariate function for each trimming site n and gene σ

$$f(n,\sigma;\boldsymbol{\beta}^{\texttt{shape}},a,b) \coloneqq \sum_{j=-a}^{b-1}\sum_{u\in\{\texttt{E},\texttt{W},\texttt{P}\}} \beta_{uj}^{\texttt{shape}}\cdot \textsf{shape}^{u}(j,\textsf{seq}_{\text{expd}}(n,\sigma,a,b)) + \sum_{d=-a+1}^{b-1}\sum_{v\in\{\texttt{R},\texttt{H}\}} \beta_{vd}^{\texttt{shape}}\cdot \textsf{shape}^{v}(d,\textsf{seq}_{\text{expd}}(n,\sigma,a,b)). \tag{D.7}$$

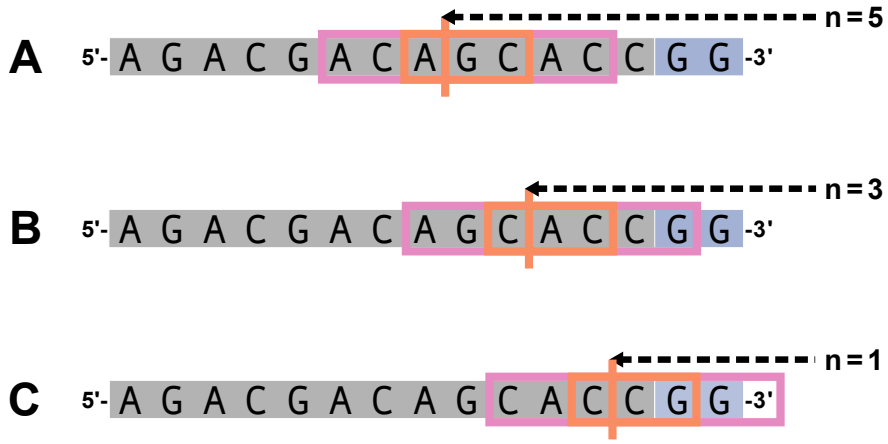

Figure D.3: Let $a = 1$ and $b = 2$. The 3 nucleotide trimming sequence window is shown in the orange box and the trimming site is shown by the vertical orange line. The 7 nucleotide expanded trimming sequence window is represented by the pink boxes in addition to the original trimming sequence window orange box. An arbitrary gene sequence is highlighted in gray and the two possible P-nucleotides are highlighted in purple. (**A**) For $n = 5$, both the 7 nucleotide expanded trimming sequence window and the original 3 nucleotide trimming sequence window will not contain P-nucleotides. (**B**) For $n = 3$, the 7 nucleotide expanded trimming sequence window will contain one P-nucleotide and the original 3 nucleotide trimming sequence window will not contain P-nucleotides. (**C**) For $n = 1$, the 7 nucleotide expanded trimming sequence window will be "incomplete" (contain less than 7 nucleotides), and thus, will be invalid for estimating DNA-shape for the nucleotides within the original trimming sequence window.

Length

We can think of the trimming amount n as a measure of the sequence-independent length from the end of the gene for each gene and trimming site, and define $\beta^{\texttt{ldist}}$ to be a *length* model coefficient. As such, we can define a length covariate function for each trimming site

n

$$f(n, \sigma; \beta^{\texttt{ldist}}) := \beta^{\texttt{ldist}} \cdot n. \tag{D.8}$$

D.2 Extended model validation methods

Calculating the expected per-sequence conditional log loss across the full V-gene training data set

With the full V-gene training set, we can train each model of interest as described above in (3.10) to obtain a trained model $\mathcal{M}$. After this model fitting, we can calculate the expected per-sequence conditional log loss of the model, $\mathcal{M}$, for the full V-gene training data set, $\mathbf{V}$, using the procedure described above in (3.11). Here, we use the full V-gene data set as both the training data set and the testing data set. Models that have lower expected per-sequence conditional log loss on the V-gene training data set will indicate that the model has a better fit. Model evaluation using held-out testing sets, as described below, is required for evaluating model generalizability.

Calculating the expected per-sequence conditional log loss across held-out samples

Because our goal is to learn a model that is gene-agnostic, we will evaluate the performance and generalizability of each model by calculating the expected per-sequence conditional log loss using many different held-out data sets. A model that is generalizable across many genes will perform well and have a good fit across all held-out samples despite their varying gene compositions. To test this, we will create each random, held-out sample from the original training data set by cluster-sampling all observations from V-gene allele groups, σ_V, uniformly at random. We will refer to each random, held-out sample as the "held-out testing set." Let G be the total number of unique V-gene allele groups in the original data set. Let $G_{\text{test}} = \text{Round}(0.3 \cdot G)$ be an integer which represents the number of unique genes included

in each "held-out testing set." As such, we can sample each gene σ_V with probability

$$P_{\text{sample}}(S = \sigma_V) \coloneqq \frac{1}{G}$$

such that the probability of each "held-out testing set" $\mathbf{H}$ is given by

$$\begin{aligned} P(\mathbf{H}) &= \prod_{\sigma_V=1}^{G_{\text{test}}} P_{\text{sample}}(S = \sigma_V) \\ &= \prod_{\sigma_V=1}^{G_{\text{test}}} \frac{1}{G}. \end{aligned} \tag{D.9}$$

The remaining genes not sampled as part of the "held-out testing set" $\mathbf{H}$ will compose the "training set" $\mathbf{T}$. Using this "training set," we can train each model of interest as described above in (3.10). After this model training, we can calculate the expected per-sequence conditional log loss of the model, $\mathcal{M}$ for the "held-out testing set", $\mathbf{H}$, as described above in (3.11). To achieve an unbiased estimate of the model performance, we will repeat the above procedure across 20 unique held-out testing sets and calculate the expected per-sequence conditional log loss across all samples. As such, the expected per-sequence conditional log loss across these random samples is given by

$$E[\ell(\mathcal{M})] = \sum_{\mathbf{H}=1}^{20} P(\mathbf{H}) \cdot \ell(\mathcal{M} \mid \mathbf{H}). \tag{D.10}$$

We use the same, unique held-out testing sets to calculate the expected per-sequence conditional log loss of each model of interest, and thus, we can compare model fit and generalizability by directly comparing the expected per-sequence conditional log loss of each model. Models that have lower expected per-sequence conditional log loss will indicate a that the model is a better fit and is more generalizable across genes.

Calculating the expected per-sequence conditional log loss across held-out samples of the "most-different" V-genes

While the previously-described procedure for evaluating the expected per-sequence conditional log loss across held-out samples of the V-gene data set provided a metric for evaluating model generalizability across different gene sets, we were interested in evaluating model performance for groups of genes which were considered "most-different" sequence-wise. Many of the germline V-gene sequences are quite similar, however, there are subgroups of these sequences which share unique sequence traits. We can characterize these "most-different" V-genes by either using only the "terminal" V-gene sequences (e.g. that last 24 nucleotides of each sequence which is directly parameterized in the models) or using the entire V-gene sequences.

To define the "most-different" V-gene allele group using the "terminal" V-gene sequences, we first calculate the pairwise hamming distance between each gene-allele group pair. We then use hierarchical clustering to cluster V-gene allele groups based on their pairwise hamming distances (Figure D.4A). The cluster that has the smallest average pairwise hamming distance within the cluster and the largest average pairwise hamming distance outside of the cluster is defined to be the "most-different" V-gene allele group cluster. To define the "most-different" V-gene allele group using the entire V-gene sequences, we first align all gene sequences using the `DECIPHER` package in R. Using these aligned sequences, we can then proceed with the same procedure as described for the "terminal" V-gene sequences to define the "most-different" V-gene allele group (Figure D.4B).

Once we have defined a cluster of the "most-different" V-gene allele groups, using either the "terminal" V-gene sequences or the full sequences, we can define a held-out testing data set $\mathbf{H}$ containing all data observations from the V-gene allele groups within this "most-different" V-gene allele group cluster. All data observations from the remaining gene allele groups that were not defined to be part of the "most-different" cluster will compose the

"training set" $\mathbf{T}$. Using this "training set," we can train each model of interest as described above in (3.10). After this model training, we can calculate the expected per-sequence conditional log loss of the model, $\mathcal{M}$ for the "held-out testing set", $\mathbf{H}$, as described above in (3.11). Models that have lower expected per-sequence conditional log loss will indicate better fit and generalizability across even the "most-different" genes. We can repeat this process for other V-gene allele group clusters (e.g. the "second-most-different" V-gene allele group cluster) as desired.

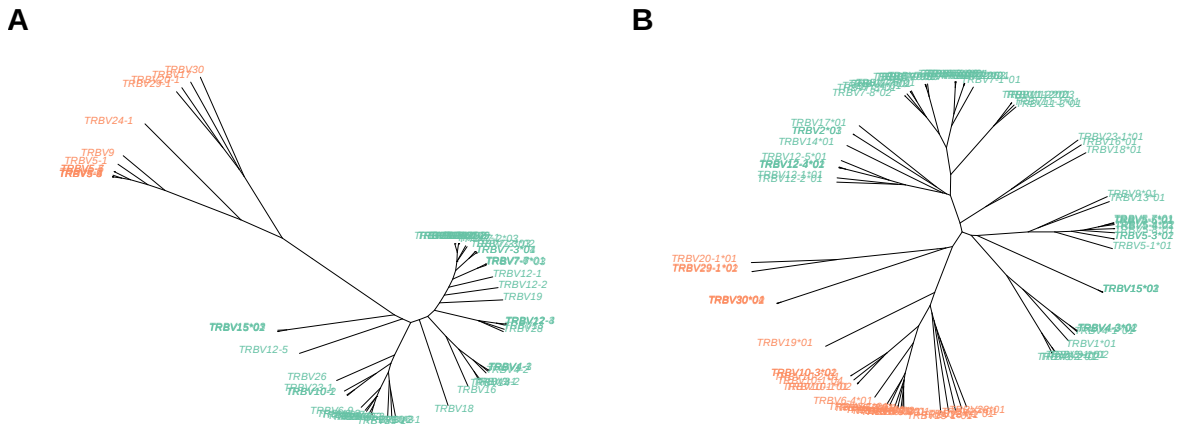

Figure D.4: Un-rooted trees of "terminal" V-gene sequences (**A**) and full-length V-gene sequences (**B**) derived from hierarchical clustering. Tips are colored according to cluster membership. The tips corresponding to the "most-different" group within each tree are colored in orange.

Calculating the expected per-sequence conditional log loss across the full J-gene data set

With the full V-gene training set, we can train each model of interest as described above in (3.10) to obtain a trained model $\mathcal{M}$. After this model fitting, we can calculate the expected per-sequence conditional log loss of the model, $\mathcal{M}$, for the full J-gene training data set, $\mathbf{J}$, using the procedure described above in (3.11). Here, we use the full V-gene data set as the training data set and the full J-gene data set as the testing data set. Models that have lower expected per-sequence conditional log loss on the J-gene data set will indicate that the

model is a better fit and is more generalizable.

Evaluating TCRβ V-gene trimming models using the expected per-sequence conditional log loss across testing data sets

To validate the performance of each model, we worked with TCRα- and TCRβ-immunosequencing data representing 150 individuals, TCRγ-immunosequencing data representing 23 individuals, and IgH-immunosequencing data representing 9 individuals from three independent validation cohorts (described above). With these data, we used the model coefficients from the previous TCRβ V-gene training run ("frozen" in git commit 093610a on our repository) and then compute the expected per-sequence conditional log loss of the model using each independent validation data set of interest. Models that have low expected per-sequence conditional log loss across all testing data sets will indicate that the model is more generalizable and less overfit to the training data. We validated each model using V- and J-gene sequences separately.

D.3 Exploring the gene-specificity of the "trimming motif"

To evaluate the specificity of the *motif* coefficients across different genes, we can compare the per-gene model predictions for the *motif* and *base-count beyond* model to a model that only contains *base-count beyond* parameters. To do this, we first use the entire V-gene data set $\mathbf{V}$ to train both the *motif* and *base-count beyond* model as before in (3.10) and a model that contains only *base-count beyond* parameters (and no motif parameters). We can then use these models to predict the probability of trimming each possible trimming amount, $2 \leq n \leq 14$, for each gene allele group sequence σ. For each of these models, we can then

calculate the per-gene root mean squared error, RMSE, for each gene σ such that

$$\text{RMSE}(\sigma, \mathcal{M}, \mathbf{V}) = \sqrt{\frac{\sum_{i=1}^{I} \sum_{n=2}^{14} \left(P_{\text{emp}}(n \mid \sigma, i) - P(n \mid \sigma; \mathcal{M}) \right)^2}{|I|}} \tag{D.11}$$

where $\mathcal{M}$ is a model trained using the V-gene training data set $\mathbf{V}$, I is the set of all individuals in the data set, $|I|$ is the length of the set of individuals I, $P_{\text{emp}}(n \mid \sigma, i)$ is the empirical conditional PDF given by (3.1) for trimming length n, gene σ, and individual $i \in I$, and $P(n \mid \sigma; \mathcal{M})$ is the predicted trimming probability from a specified model $\mathcal{M}$. We can then compare this per-gene root mean squared error for the model trained using both *motif* and *base-count beyond* parameters with a model trained using just *base-count beyond* parameters.

D.4 Sensitivity analysis for hairpin nick position

For our modeling, we assume that the initial hairpin nick occurs at the +2 position and will create two P-nucleotides at the end of the 5'-to-3' gene sequence. Assuming a different hairpin nick position would incorporate a different number of P-nucleotides at the end of the gene sequence (Figure D.5). While the hairpins are assumed to be nicked at the +2 position most frequently [29, 31], we wanted to test the sensitivity of our models to this hairpin-nick-position assumption. To do this, we assumed each of the other possible hairpin opening positions (e.g. -2, -1, 0, +1, +3) one-at-a-time and appended the appropriate number of associated of P-nucleotides given the assumed hairpin-nick-position to the 3'-end of each V-gene allele group sequence in the data set. With each of these hairpin-position data sets, we re-trained the *motif* and *base-count beyond* model as before in (3.10) and calculate the expected per-sequence conditional log loss of the model using (3.11). We can compare these expected per-sequence conditional log losses to evaluate the sensitivity of the model to the +2 hairpin nick assumption.

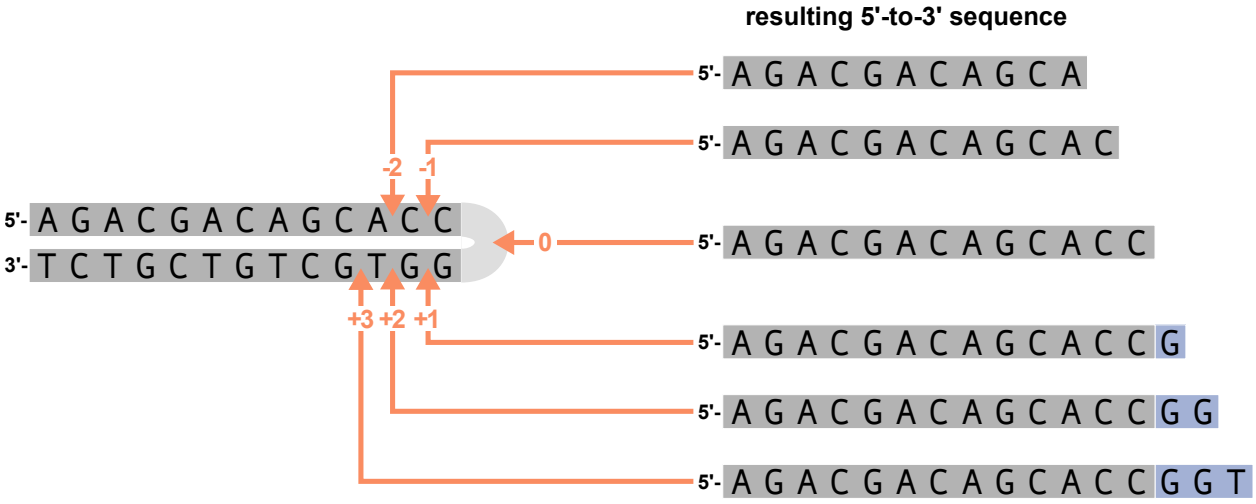

Figure D.5: An arbitrary DNA hairpin can be nicked opened at various positions near the hairpin (left figure). Hairpin nick position 0 refers to a nick at the tip of the hairpin, position -1 refers to a nick before the last nucleotide on the 5' strand, position +1 refers to a nick before the last nucleotide on the 3' strand, etc. The resulting 5'-to-3' sequences from the various nick positions for the arbitrary gene sequence are shown on the right. Nucleotides originating from the 5' strand of the DNA hairpin are highlighted in gray and P-nucleotides (originating from the 3' strand of the DNA hairpin) are highlighted in purple. The various hairpin nick positions lead to 5'-to-3' sequences that contain different amounts of P-nucleotides. Hairpin nick positions > 0 lead to 5'-to-3' sequences that contain P-nucleotides, nick positions equal to zero lead to 5'-to-3' sequences without P-nucleotides, and nick positions < 0 lead to 5'-to-3' sequences without P-nucleotides and with portions of the original 5' DNA hairpin strand removed.

D.5 Evaluating the weight of the 1x2 motif *and* two-side base-count beyond *model terms across data sets*

For each testing data set, we can measure the weight of the *1x2 motif* and *two-side base-count beyond* model terms within the full *1x2 motif + two-side base-count beyond* model. Recall that we trained the full *1x2 motif + two-side base-count beyond* model using the model covariate function given by (3.2)

$$f(n, \sigma_V; \boldsymbol{\beta}^{\texttt{motif}}, \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \coloneqq f(n, \sigma_V; \boldsymbol{\beta}^{\texttt{motif}}, a, b) + f(n, \sigma_V; \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c)$$

where n represents the number of trimmed nucleotides, σ_V represents the V-gene allele group sequence, $\boldsymbol{\beta}^{\texttt{motif}}$ represents *motif*-specific parameter coefficients, $\boldsymbol{\beta}^{\texttt{AT}}$ and $\boldsymbol{\beta}^{\texttt{GC}}$ represent *base-count-beyond*-specific parameter coefficients, a and b are non-negative integer values that represent the number of nucleotides 5' and 3' of the trimming site, respectively, that are included in the "trimming motif", and c represents the number of nucleotides 5' of the trimming site to be included in the base-count. As such, for each training data set, we can use the inferred coefficients, $\boldsymbol{\beta}^{\hat{\texttt{motif}}}$, $\boldsymbol{\beta}^{\hat{\texttt{AT}}}$, and $\boldsymbol{\beta}^{\hat{\texttt{GC}}}$, from a previous training run and define a new two-parameter model containing a scale coefficient for the *1x2 motif* terms and a second scale coefficient for the *two-side base-count beyond* terms. The covariate function for this new model is given by

$$\begin{aligned} &f(n, \sigma_V; \boldsymbol{\beta}^{\hat{\texttt{motif}}}, \boldsymbol{\beta}^{\hat{\texttt{AT}}}, \boldsymbol{\beta}^{\hat{\texttt{GC}}}, \alpha_{\texttt{motif}}, \alpha_{\texttt{count}}, a, b, c) \\ &\qquad := \alpha_{\texttt{motif}} \cdot f(n, \sigma_V; \boldsymbol{\beta}^{\texttt{motif}}, a, b) + \alpha_{\texttt{count}} \cdot f(n, \sigma_V; \boldsymbol{\beta}^{\texttt{AT}}, \boldsymbol{\beta}^{\texttt{GC}}, a, b, c) \end{aligned}$$

where $\alpha_{\texttt{motif}}$ is the scale coefficient for the *1x2 motif* terms and $\alpha_{\texttt{count}}$ is the scale coefficient for the *two-side base-count beyond* terms. We can then train this new model as described previously for each data set of interest and compare the inferred scale coefficients.

Appendix E

SUPPLEMENTARY TABLES AND FIGURES FOR CHAPTER 4

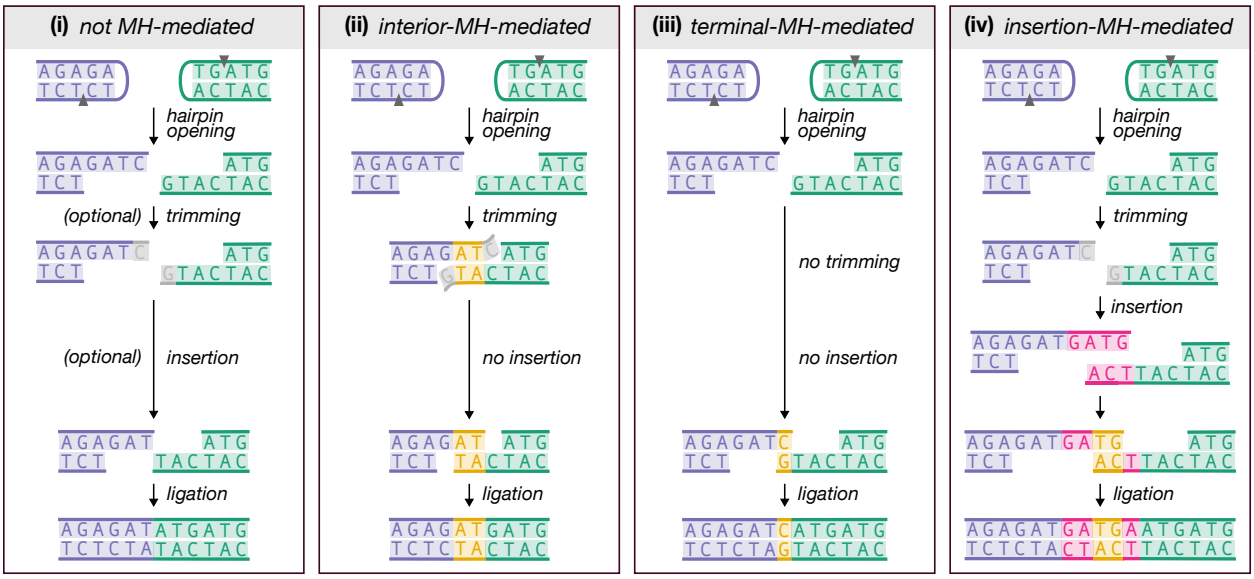

Figure E.1: Illustration of how microhomologous nucleotides could influence trimming and/or ligation during V(D)J recombination. The example depicts microhomologous regions (yellow), trimmed nucleotides (gray), and inserted nucleotides (pink) for a V-gene (purple) and J-gene (green), highlighting four possible regimes: (i) no microhomology influence, (ii) influence of interior microhomology, (iii) influence of terminal microhomology, and (iv) influence of insertion-dependent microhomology. Terminal and interior microhomologous nucleotides are germline-encoded, whereas insertion-dependent microhomologous nucleotides are randomly added by terminal deoxynucleotidyl transferase (TdT) and are not encoded in the germline. Germline-encoded microhomologous regions are classified as terminal microhomology when they are located at the ends of untrimmed sequences, and as interior microhomology when they are located within the interior regions of the gene sequences.

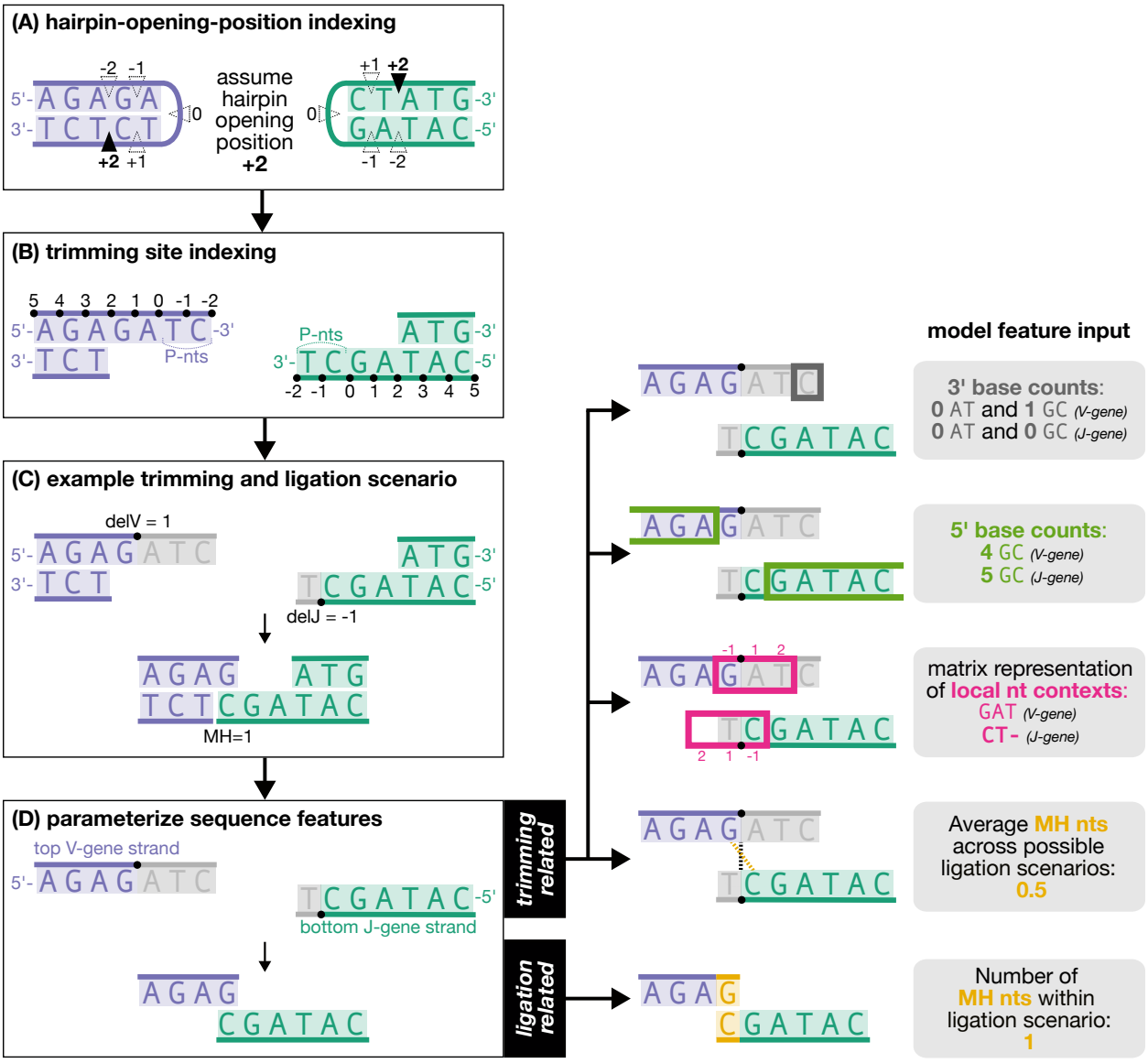
(A) hairpin-opening-position indexing
assume hairpin opening position +2
(B) trimming site indexing
P-nts
(C) example trimming and ligation scenario
delV = 1
delJ = -1
MH=1
(D) parameterize sequence features
top V-gene strand
bottom J-gene strand
trimming related
ligation related
model feature input
3' base counts:
0 AT and 1 GC (V-gene)
0 AT and 0 GC (J-gene)
5' base counts:
4 GC (V-gene)
5 GC (J-gene)
matrix representation of local nt contexts:
GAT (V-gene)
CT- (J-gene)
Average MH nts across possible ligation scenarios:
0.5
Number of MH nts within ligation scenario:
1

Figure E.2: Overview of how sequences are transformed into features for regression. **(A)** During early stages of V(D)J recombination, the hairpin of each gene is opened. Hairpins are most frequently opened at position +2, but other positions are possible. For modeling, we assume all hairpins open at +2. **(B)** Hairpin opening creates a 4-nucleotide single-stranded overhang, with 2 nucleotides considered P-nucleotides. Each gene can then undergo nucleotide trimming. Trimming sites are indexed such that negative values represent P-nucleotide deletions, while positive values represent coding sequence deletions. For example, a deletion of 0 trims to the end of the germline gene (removing 2 P-nucleotides), and -2 indicates no trimming of P-nucleotides or gene sequence. This indexing is consistent with the IGoR software [11]. **(C)** In this example trimming and ligation scenario, the V-gene is trimmed by 3 nucleotides (V-trimming site $\mathrm{delV} = 1$), and the J-gene is trimmed by 1 nucleotide (J-trimming site, $\mathrm{delJ} = -1$). The trimmed genes are ligated with 1 nucleotide of microhomology. **(D)** Illustration of sequence features and their alignment for an arbitrary V- and J-gene pair and example trimming and ligation scenario. For modeling, only the top strand of the V-gene and the bottom strand of the J-gene are considered, consistent with the most common overhang polarities. Features related to ligation include the microhomology-related feature (yellow), which captures the number of microhomologous nucleotides in the ligation scenario. Features related to trimming include the microhomology-related feature (yellow), which captures the average number of microhomologous nucleotides across all possible ligation scenarios for the given trimming scenario; the trimming motif features (pink), which represent the identities of nucleotides adjacent to the trimming site and are indexed relative to the site, with negative indices indicating positions 5' of the site and positive indices indicating positions 3'; the 5' base count parameters (green), which capture the GC nucleotide count within 10 nucleotides upstream of the trimming motif; and the 3' base count parameters (gray), which capture the GC and AT nucleotide counts downstream of the trimming motif, if applicable.

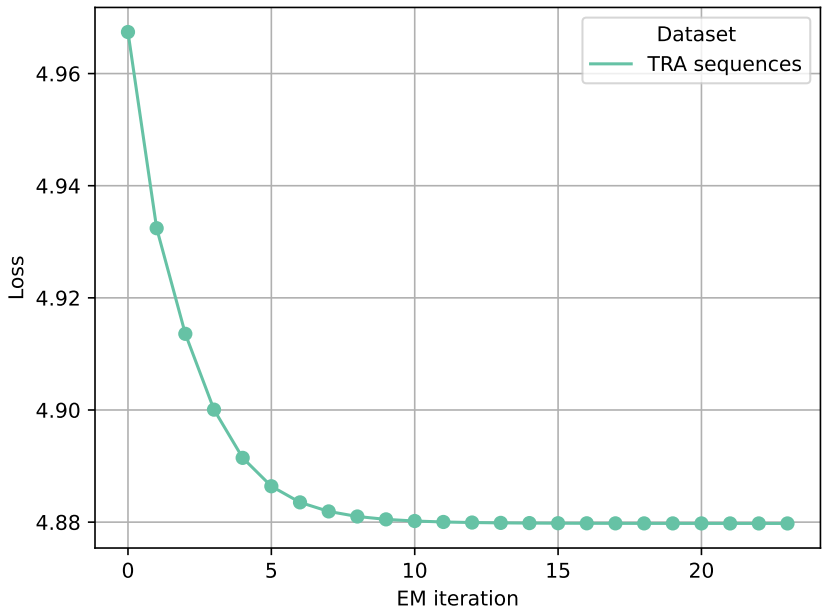

Figure E.3: Convergence of the expectation-maximization (EM) algorithm using a training dataset of non-productive TCRα sequences without N-insertions, alongside their corresponding sets of potential microhomology-adapted annotations (as detailed in Methods). The y-axis represents the expected per-sequence log loss, as defined in $(\mathrm{F}.20)$.

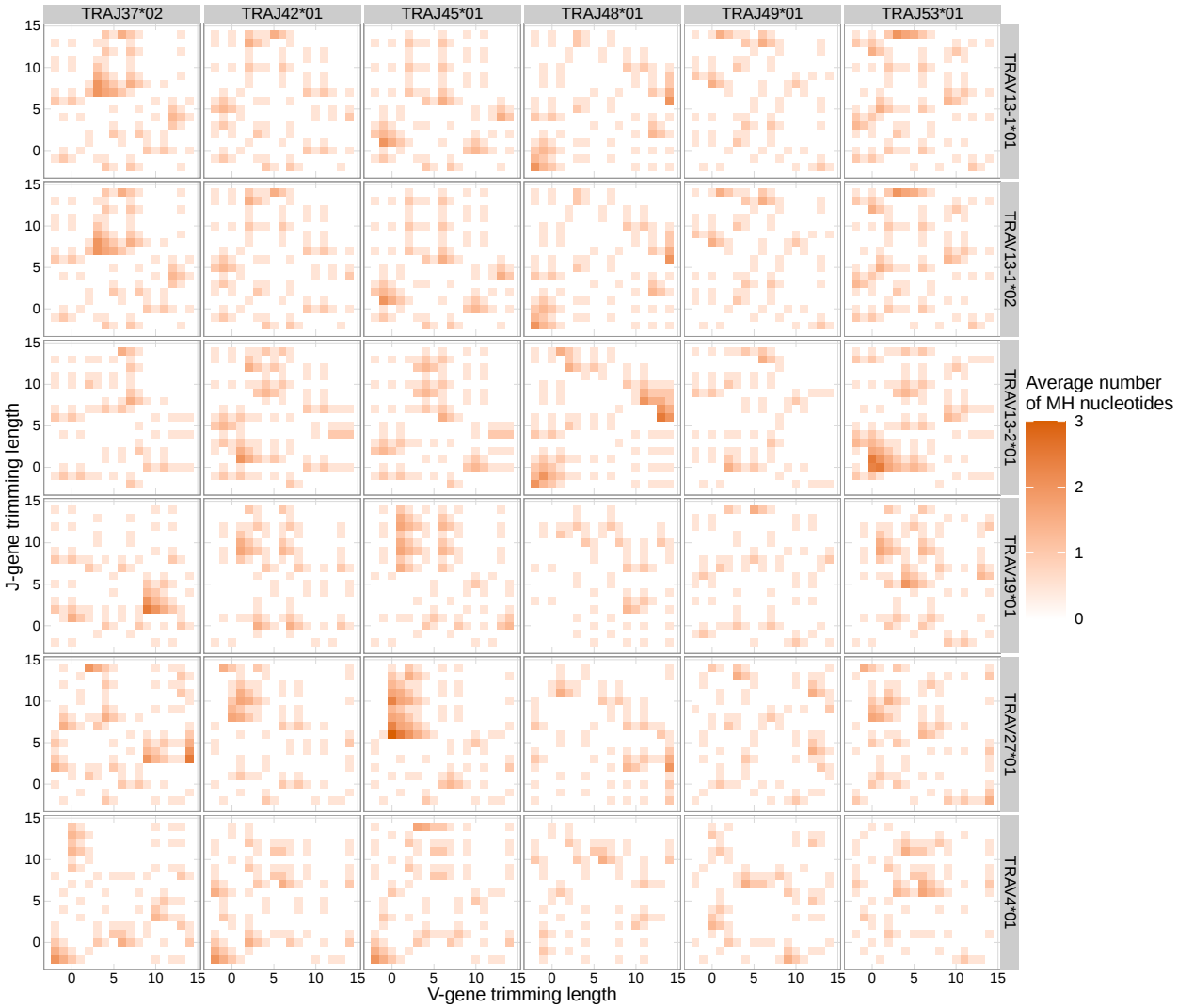

Figure E.4: The distribution of complementary sequence regions capable of forming microhomologous regions during V(D)J recombination varies by trimming amounts and V-J gene pairs. Depending on the gene pair, there is potential for both interior and terminal microhomology (MH). For instance, the TRAV13-1*01 and TRAJ48*01 gene pair shows potential for terminal microhomology (e.g. the average number of MH nucleotides for the untrimmed sequences—both genes trimmed at the -2 site—is nonzero), as well as interior microhomology. In contrast, the TRAV13-1*01 and TRAJ37*02 gene pair lacks terminal microhomology potential (e.g. the average number of MH nucleotides for the untrimmed sequences is zero) but has an abundance of interior microhomology. The average microhomology counts are calculated across all possible ligation scenarios for each gene pair trimming scenario. Only the most frequently used gene pairs are plotted here.

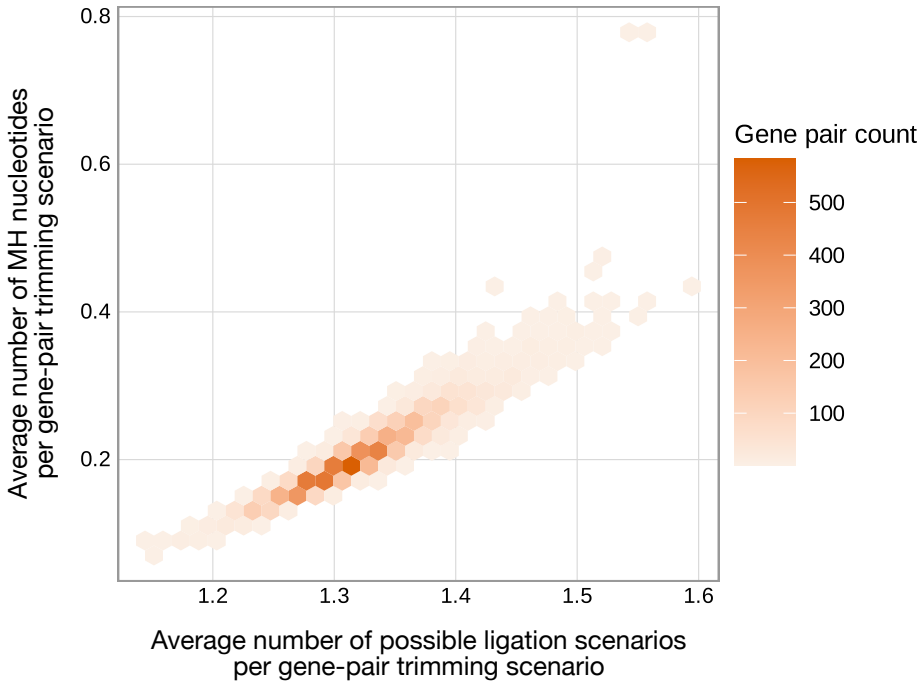

Figure E.5: Complementary sequence regions capable of forming microhomologous regions during V(D)J recombination are common between germline V- and J-genes in the *TRA* locus. As the average number of microhomologous nucleotides increases for a given gene pair trimming scenario, so does the average number of possible ligation scenarios. The median average number of microhomologous nucleotides across all possible gene pair trimming scenarios is 0.1978, corresponding to a median of 1.3149 possible ligation scenarios. Since a median of exactly one ligation scenario would indicate that all scenarios involve zero microhomology, this suggests that most trimming scenarios result in multiple ligation outcomes—both with and without microhomology. The average values are calculated across all trimming scenarios for each gene pair, with each gene pair plotted only once.

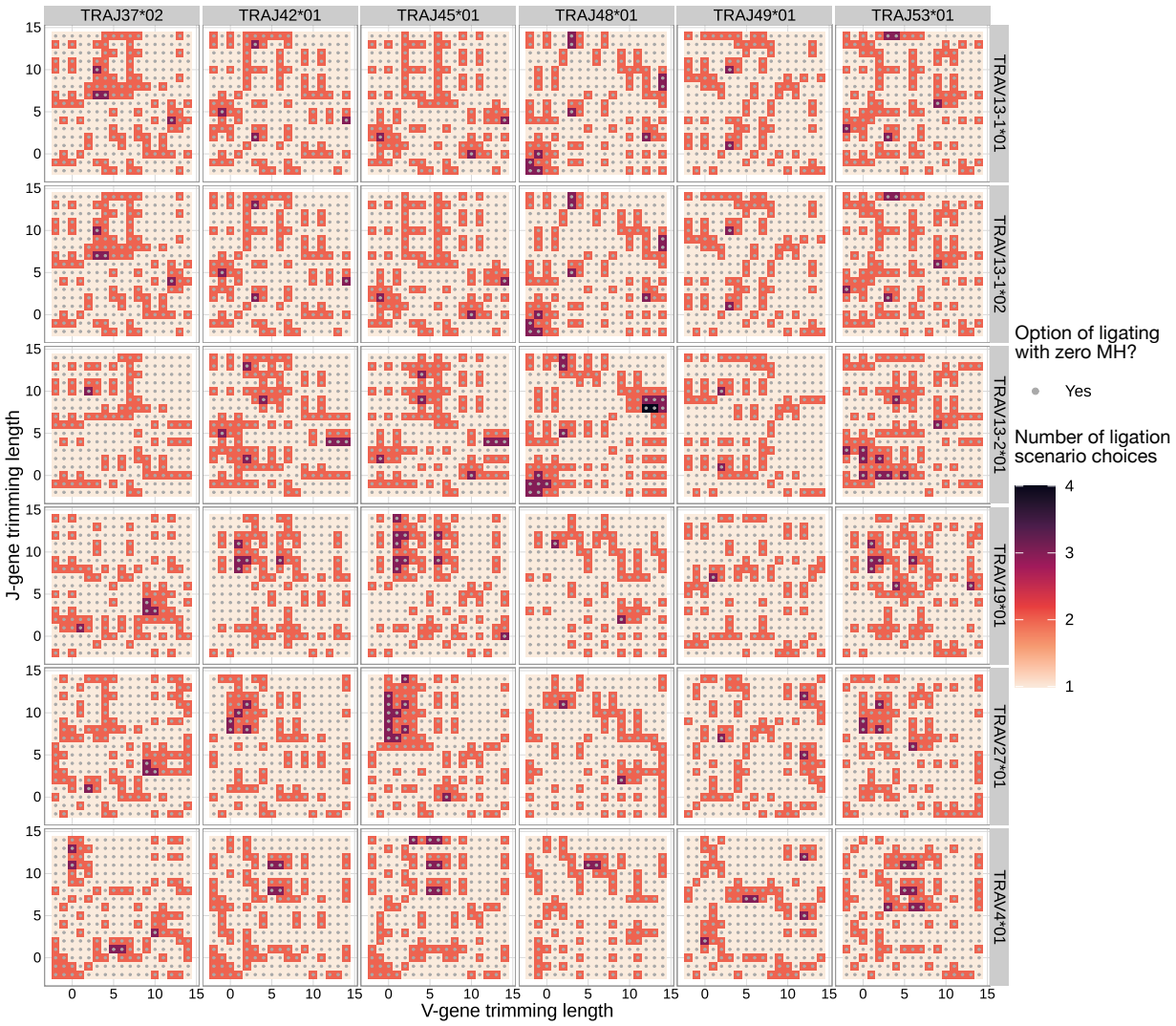

Figure E.6: The distribution of trimming scenarios with multiple ligation options (e.g. varying amounts of microhomology in the observed sequence) varies across V-J gene pairs. All gene pair trimming scenarios can ligate with zero nucleotides of microhomology (indicated by gray dots in the plot), providing at least one ligation scenario choice. Depending on the gene pair, there may be potential for both interior- and terminal-microhomology-mediated trimming and ligation (e.g. ligation using nonzero microhomology), which would increase the number of possible ligation scenario choices. For example, the TRAV13-1*01 and TRAJ48*01 gene pair shows potential for terminal-microhomology-mediated ligation (e.g. ligating the untrimmed sequences—both genes trimmed at the -2 site—using nonzero microhomology), as well as interior-microhomology-mediated ligation. In contrast, the TRAV13-1*01 and TRAJ37*02 gene pair lacks potential for terminal-microhomology-mediated ligation (e.g. the untrimmed sequences can only be ligated using zero microhomology) but has substantial potential for interior-microhomology-mediated ligation. Only the most frequently used gene pairs are plotted here.

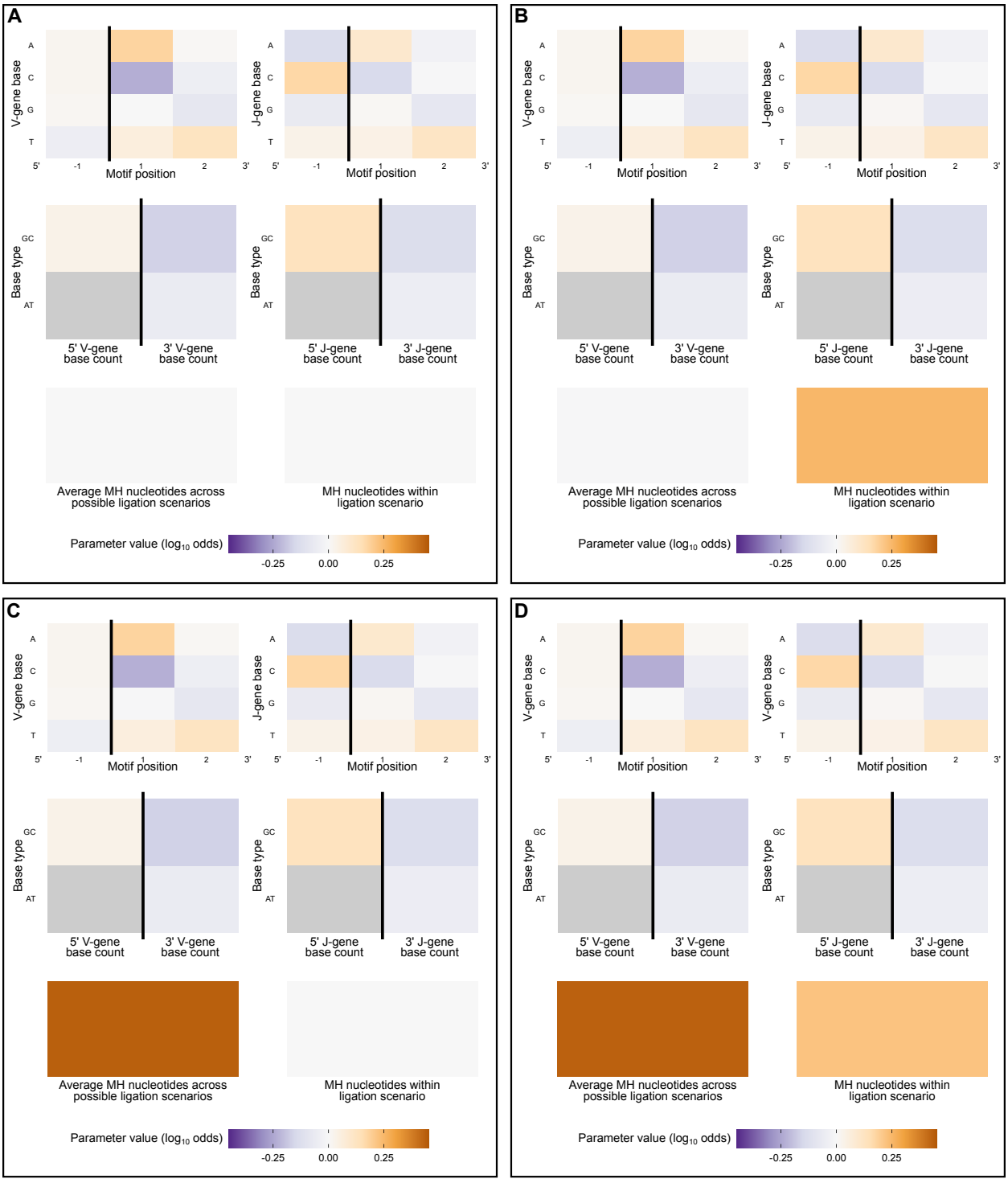
A
V-gene base
J-gene base
A
C
G
T
5'
-1
1
2
3'
Motif position
Base type
GC
AT
5' V-gene base count
3' V-gene base count
5' J-gene base count
3' J-gene base count
Average MH nucleotides across possible ligation scenarios
MH nucleotides within ligation scenario
Parameter value (log10 odds)
-0.25
0.00
0.25
B
C
D

Figure E.7: Parameters inferred from simulated data emulating varying levels of microhomology involvement in V(D)J recombination trimming and ligation (see Appendix F). The model was trained using four different simulated datasets. (**A**) Parameters inferred from data where microhomology does not influence trimming or ligation choices. In this scenario, the microhomology-related parameters show no signal. (**B**) Parameters inferred from data where microhomology increases ligation probabilities but does not affect trimming probabilities. Here, the trimming-related microhomology parameter shows no signal, while the ligation-related parameter shows a strong positive signal. (**C**) Parameters inferred from data where microhomology increases trimming probabilities but does not affect ligation probabilities. In this case, the ligation-related microhomology parameter shows no signal, while the trimming-related parameter shows a strong positive signal. (**D**) Parameters inferred from data where microhomology increases both trimming and ligation probabilities. As expected, both microhomology-related parameters exhibit strong positive signals. The patterns observed in this simulated dataset closely matches those inferred from actual data. All trimming motif and two-side base count parameters remain consistent across all simulated datasets.

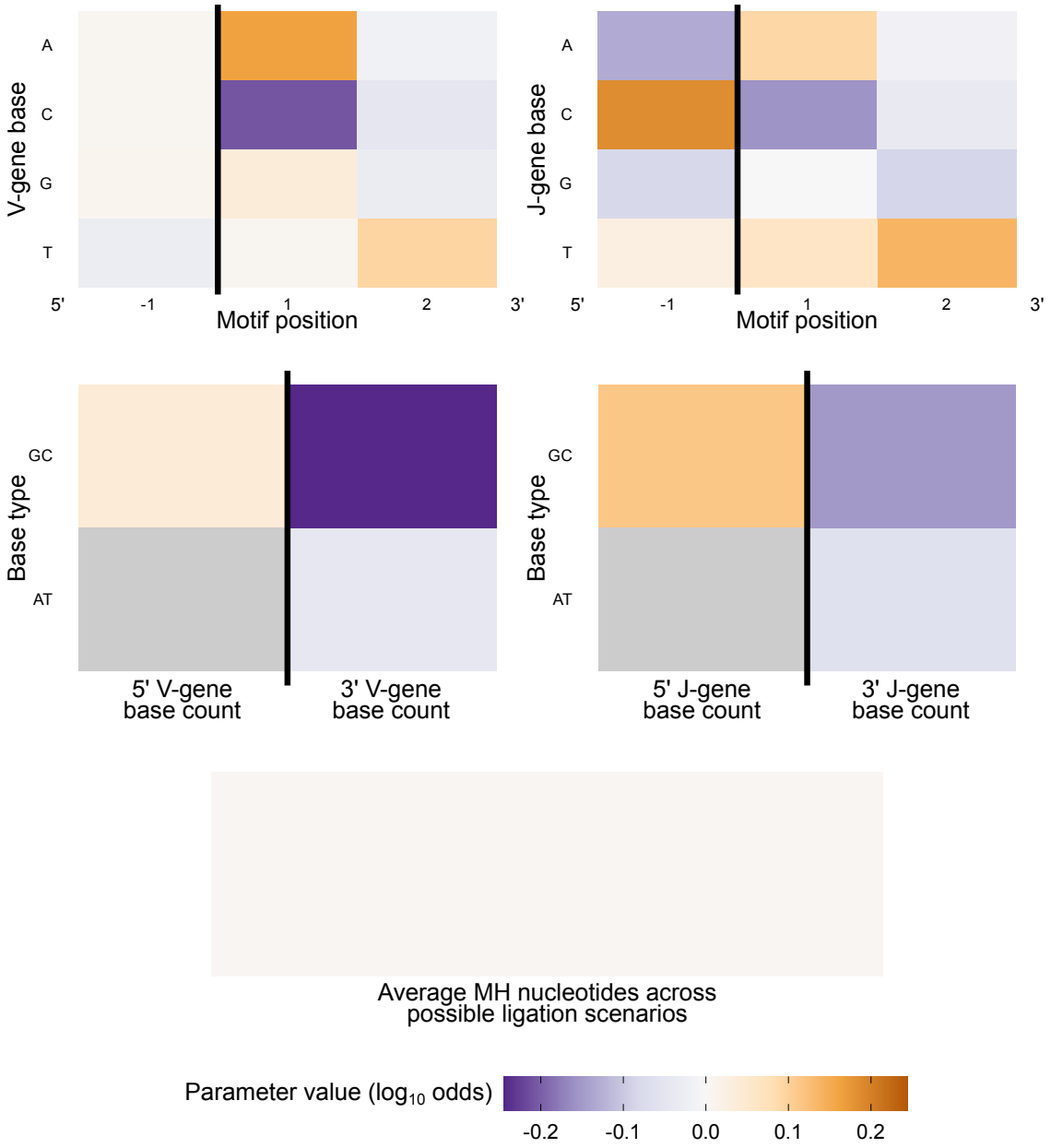

Figure E.8: Parameters inferred from a model trained on sequences with N-insertions, which lack germline-dependent ligation. Due to the unknown composition of inserted nucleotides prior to ligation, ligation patterns cannot be detected, allowing this model to specifically explore trimming probabilities independent of ligation (see Appendix F). The ligation-related microhomology parameter was excluded from this model. Inferred trimming motif and two-side base count parameters align with previous analyses of individual V- and J-gene sequences [3]. The inferred trimming-related microhomology parameter indicates that microhomology has only a minimal effect on trimming probabilities, suggesting a limited independent role of microhomology in trimming.

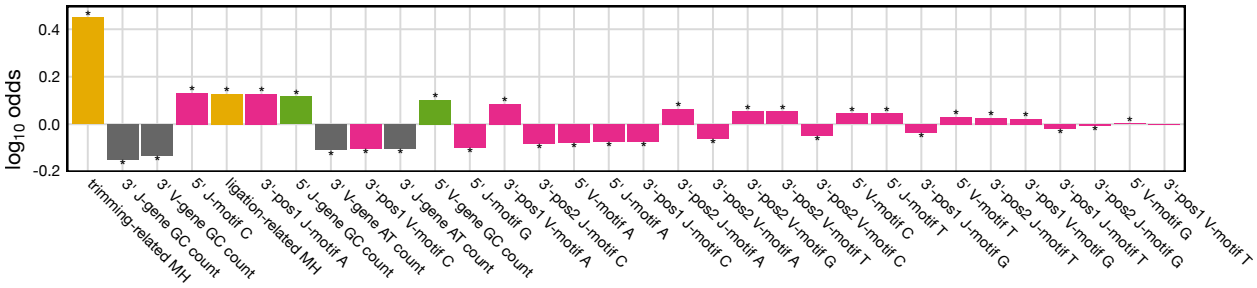

Figure E.9: Both MH-related parameters (gold) exhibit large effect sizes, with the trimming-related MH parameter showing the strongest positive effect. Stars indicate parameters significant at a Bonferroni-corrected threshold of 0.0016.

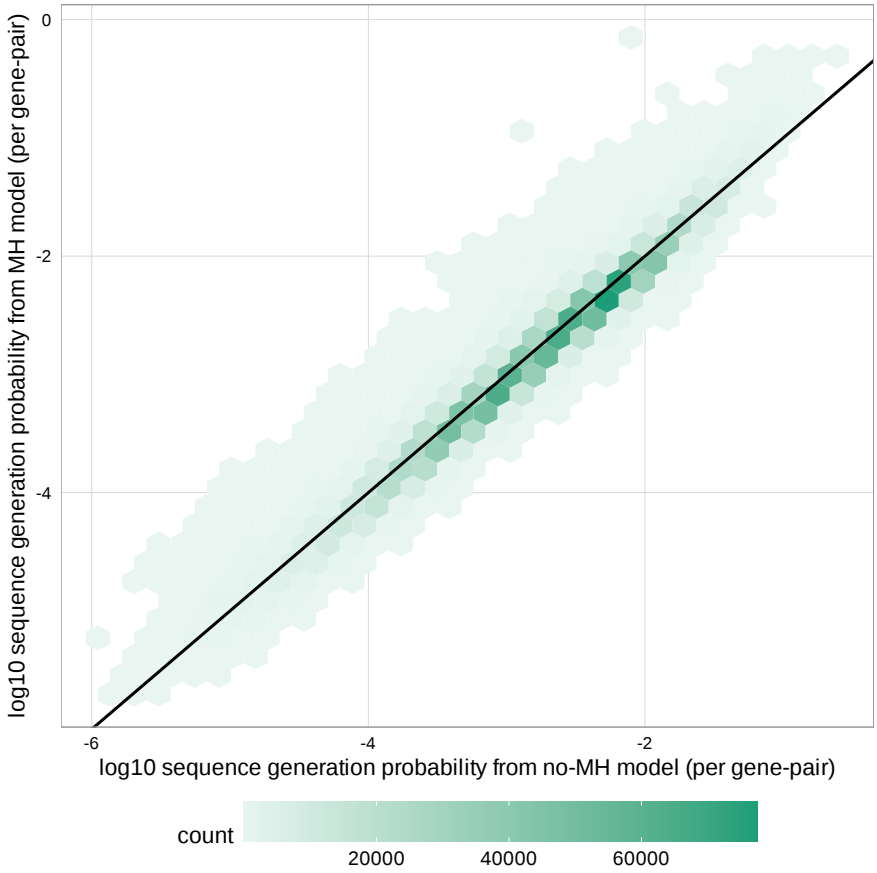

Figure E.10: Sequence generation probabilities differ slightly between models that include microhomology effects and those that do not. Both the model parameterizing microhomology (MH model) and the one that does not (no-MH model) include trimming motif and base-count parameters and are identical except for the inclusion of microhomology terms. For each model, we calculate sequence generation probabilities as the aggregated probability of all possible trimming and ligation scenarios for a sequence, normalized across all sequences for a given V-J pair.

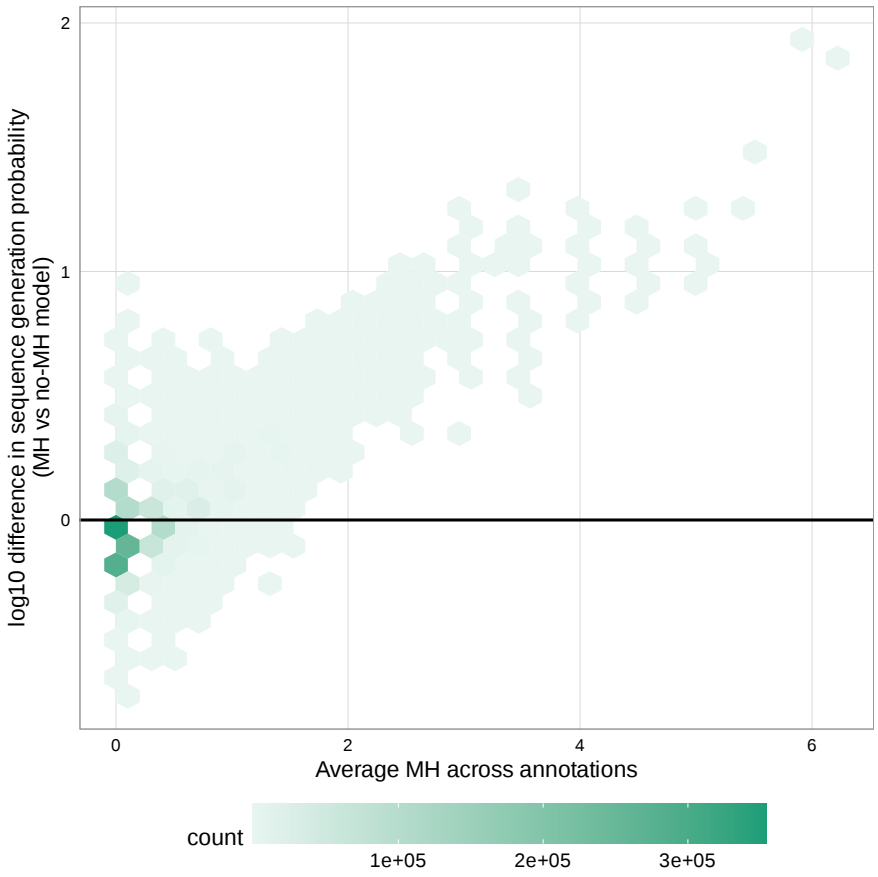

Figure E.11: As the average number of microhomologous nucleotides across possible annotation scenarios for a given sequence increases, the difference in sequence generation probabilities between the model parameterizing microhomology (MH model) and the one that does not (no-MH model) becomes larger. Both models include trimming motif and base-count parameters and are identical except for the inclusion of microhomology terms. The plotted differences are calculated using probabilities normalized across all possible sequences for each V-J gene pair, ensuring the sum of these differences is zero.

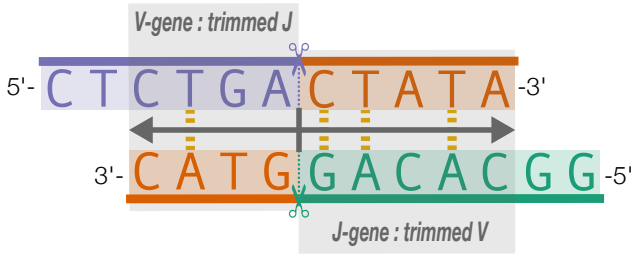

Figure E.12: Cartoon showing the alignment of V-gene (purple) and J-gene (green) sequences at their inferred trimming sites (marked with scissors) without gaps. Trimmed regions of each sequence are shown in orange. We focus on two regions: (1) overlap between V-gene and trimmed J-gene (*V-gene:trimmed-J*) and (2) overlap between J-gene and trimmed V-gene (*J-gene:trimmed-V*). The function h in $(\mathrm{F}.1)$ counts contiguous, complementary nucleotides from the aligned trimming site (indexed as zero). Arrows show the counting direction for each region. Contiguous, complementary nucleotides are counted only if adjacent to the trimming site. In this example, *V-gene:trimmed-J* has 0 contiguous complementary nucleotides, and *J-gene:trimmed-V* has 2.

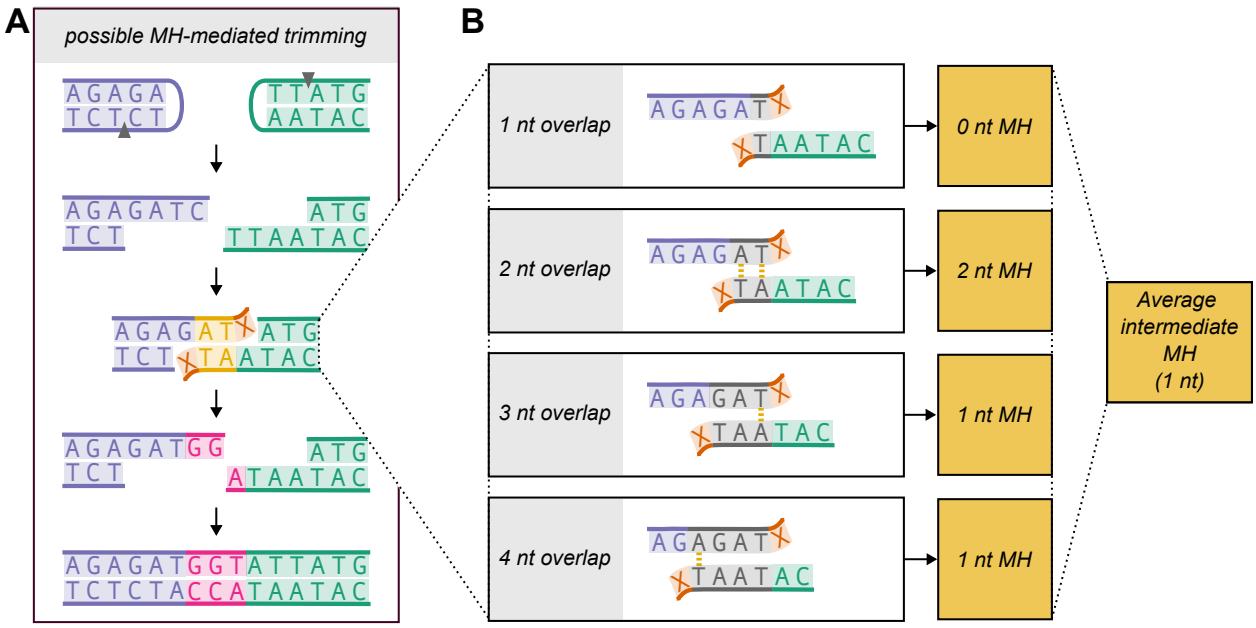

Figure E.13: (**A**) Diagram illustrating V(D)J recombination steps, emphasizing possible internal/intermediate microhomology during trimming. This intermediate microhomology may occur in both single-stranded overhangs and double-stranded regions due to sequence breathing. The final joined sequence, post-N-insertion (pink), is shown in the last box. (**B**) Definition of overlapping sequence regions for specified V-gene, J-gene, and trimming scenario. The top strand of a V-gene and the bottom strand of a J-gene are each shown with one nucleotide removed (trimmed nucleotides in orange). Overlapping regions (highlighted in gray) are obtained by aligning the sequences such that an integer value, a, of nucleotides 5' of each trimming site overlap. The value of a ranges from 1 to 4 nucleotides. Despite what is shown in this example, two genes can be trimmed by different amounts and still yield these overlapping regions. Complementary nucleotides in these regions are indicated by vertical yellow lines. The final joined sequence post-trimming and insertion is shown in the last box of panel (A).

Table E.1: Summary of all notation used in our modeling.

Variable	Description
General notation	
X	arbitrary sampled sequence
$\mathcal{X}$	set of sampled sequences
V, J	random variables for the V- and J-gene
VJ	ordered pair of genes, (V, J)
I	random variable for number of N-insertions
Q	deterministic variable for sequence productivity
MH	random variable for number of microhomologous nucleotides within the observed sequence; also referred to as a "ligation scenario"
$\mathrm{delV}, \mathrm{delJ}$	random variables for trimming amounts from V/J-gene
delVJ	pair of trimming amounts, $(\mathrm{delV}, \mathrm{delJ})$; also referred to as a "trimming scenario"
$\mathrm{delV}_i, \mathrm{delJ}_i$	random variables for nucleotides deleted from the V/J-gene, inferred directly by IGoR
A_X	set of possible trimming and ligation annotations for a sequence X
Motif parameter notation	
$\boldsymbol{\beta}_V^{\mathtt{motif}}, \boldsymbol{\beta}_J^{\mathtt{motif}}$	set of all V/J-gene motif parameters, defined in detail within Appendix F
$f_{\mathtt{motif}}(\mathrm{delV}, \mathrm{V}; \boldsymbol{\beta}_V^{\mathtt{motif}})$	V-gene motif weight function (F.5)
$f_{\mathtt{motif}}(\mathrm{delJ}, \mathrm{J}; \boldsymbol{\beta}_J^{\mathtt{motif}})$	J-gene motif weight function (F.5)
Base count parameter notation	
$\boldsymbol{\beta}_V^{\mathtt{AT}}$, $\boldsymbol{\beta}_V^{\mathtt{GC}}$	V-gene AT/GC base count parameters, defined in detail within Appendix F
$\boldsymbol{\beta}_J^{\mathtt{AT}}$, $\boldsymbol{\beta}_J^{\mathtt{GC}}$	J-gene AT/GC base count parameters
$f_{\mathtt{count}}(\mathrm{delV}, \mathrm{V}; \boldsymbol{\beta}_V^{\mathtt{AT}}, \boldsymbol{\beta}_V^{\mathtt{GC}})$	V-gene base count weight function (F.10)
$f_{\mathtt{count}}(\mathrm{delJ}, \mathrm{J}; \boldsymbol{\beta}_J^{\mathtt{AT}}, \boldsymbol{\beta}_J^{\mathtt{GC}})$	J-gene base count weight function (F.10)

Microhomology parameter notation	
$\beta^{\texttt{trimMH}}, \beta^{\texttt{ligMH}}$	trimming/ligation microhomology parameters
$f_{\texttt{trimMH}}(\text{delVJ}, \text{VJ}; \beta^{\texttt{trimMH}})$	trimming-related microhomology weight function (F.11)
$f_{\texttt{ligMH}}(\text{MH}; \beta^{\texttt{ligMH}})$	ligation-related microhomology weight function (F.12)
Model notation	
$\boldsymbol{\beta}_{\texttt{trim}}$	set of all trimming-related model parameters: $\boldsymbol{\beta}_V^{\texttt{motif}}, \boldsymbol{\beta}_J^{\texttt{motif}}, \boldsymbol{\beta}_V^{\texttt{AT}}, \boldsymbol{\beta}_J^{\texttt{AT}}, \boldsymbol{\beta}_V^{\texttt{GC}}, \boldsymbol{\beta}_J^{\texttt{GC}}, \beta^{\texttt{trimMH}}$
$\boldsymbol{\beta}_{\texttt{lig}}$	set of all ligation-related model parameters: $\beta^{\texttt{ligMH}}$
$f_{\texttt{trim}}(\text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{trim}})$	trimming-related weight function (F.2)
$f_{\texttt{lig}}(\text{delVJ}, \text{MH}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})$	ligation-related weight function (F.3)
$P(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}})$	two-step conditional logit model (F.17)
$P_{\text{annot}}(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0; X, \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}})$	model-derived trimming and ligation annotation probability (F.18)
$\ell(\boldsymbol{\beta}'_{\texttt{lig}}, \boldsymbol{\beta}'_{\texttt{trim}} \mid \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}}; X, \text{Q}, \text{I} = 0)$	expected log-likelihood for single sequence (F.19)
$\mathcal{L}(\boldsymbol{\beta}'_{\texttt{lig}}, \boldsymbol{\beta}'_{\texttt{trim}} \mid \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}}; \mathcal{X}, \text{Q}, \text{I} = 0)$	log-likelihood for observed sequences (F.20)

Table E.2: Summary of all parameters and parameter-specific weights for an arbitrary gene pair $\mathrm{VJ} = (\mathrm{V}, \mathrm{J})$ and trimming scenario $\mathrm{delVJ} = (\mathrm{delV}, \mathrm{delJ})$. Detailed definitions of each parameter and corresponding weights are located within Appendix F.

Parameter	**Description**	**Notation**	**Parameter weight**
V-gene motif parameters	parameterizes the importance of several nucleotides on either side of the V-gene trimming site	$\boldsymbol{\beta}_V^{\texttt{motif}}$	$f_{\texttt{motif}}(\mathrm{delV}, \mathrm{V}; \boldsymbol{\beta}_V^{\texttt{motif}})$ (F.5)
J-gene motif parameters	parameterizes the importance of several nucleotides on either side of the J-gene trimming site	$\boldsymbol{\beta}_J^{\texttt{motif}}$	$f_{\texttt{motif}}(\mathrm{delJ}, \mathrm{J}; \boldsymbol{\beta}_J^{\texttt{motif}})$ (F.5)
V-gene base count parameters	parameterizes the importance of the counts of GC and AT nucleotides beyond the V-gene trimming motif	$\boldsymbol{\beta}_V^{\texttt{AT}}$ and $\boldsymbol{\beta}_V^{\texttt{GC}}$	$f_{\texttt{count}}(\mathrm{delV}, \mathrm{V}; \boldsymbol{\beta}_V^{\texttt{AT}}, \boldsymbol{\beta}_V^{\texttt{GC}})$ (F.10)
J-gene base count parameters	parameterizes the importance of the counts of GC and AT nucleotides beyond the J-gene trimming motif	$\boldsymbol{\beta}_J^{\texttt{AT}}$ and $\boldsymbol{\beta}_J^{\texttt{GC}}$	$f_{\texttt{count}}(\mathrm{delJ}, \mathrm{J}; \boldsymbol{\beta}_J^{\texttt{AT}}, \boldsymbol{\beta}_J^{\texttt{GC}})$ (F.10)
inter-mediate microhomol-ogy parameters	parameterizes the importance of the average number of non-contiguous intermediate microhomology between a gene pair given a trimming scenario	$\beta^{\texttt{iMH}}$	$f_{\texttt{iMH}}(\mathrm{delV}, \mathrm{delJ}, \mathrm{V}, \mathrm{J}; \beta^{\texttt{iMH}})$ (F.29)
trimming-related observed mi-crohomology parameters	parameterizes the importance of the average number of contiguous microhomology across all possible ligation scenarios for a given trimming scenario and gene pair	$\beta^{\texttt{trimMH}}$	$f_{\texttt{trimMH}}(\mathrm{delVJ}, \mathrm{VJ}; \beta^{\texttt{trimMH}})$ (F.11)
ligation-related observed mi-crohomology parameters	parameterizes the importance of the number of contiguous microhomologous nucleotides for a given ligation scenario, trimming scenario, and gene pair	$\beta^{\texttt{ligMH}}$	$f_{\texttt{ligMH}}(\mathrm{MH}; \beta^{\texttt{ligMH}})$ (F.12)

Appendix F

SUPPLEMENTARY METHODS FOR CHAPTER 4

F.1 Extended dataset descriptions

TCRα training dataset

We downloaded TCRα repertoire sequence data from thymocyte samples of 10 immunologically healthy infants (aged 0-1 years) from the Adaptive Biotechnologies immuneACCESS database, following links provided in the original publications [118, 119].

Initial V(D)J recombination annotations were assigned to each sequence using the IGoR software (version 1.4.0) [11], which generates potential recombination annotations alongside their corresponding likelihoods. For each sequence, we identified the ten highest-probability recombination annotations and sampled one annotation based on posterior probabilities to assign an initial annotation. This included V- and J-gene assignments, trimming lengths, and the number of N-insertions.

Using these initial annotations, we filtered out sequences with inferred N-insertions (as determined by IGoR) to focus on potential germline microhomology-mediated ligation events. Sequences with N-insertions were excluded because their presence likely indicates that ligation did not involve germline microhomology. For the remaining sequences, we processed the initial annotations to determine all possible trimming and ligation scenarios

based on the IGoR-inferred V- and J-gene assignments and N-insertion amounts. Since IGoR does not explicitly account for microhomology and assigns shared nucleotides to only one gene segment, we did not directly use the IGoR-inferred trimming annotations. Instead, we adapted the trimming scenario annotations to account for germline-encoded microhomology, generating a set of possible trimming and ligation scenarios for each sequence, including scenarios involving germline-encoded microhomologous nucleotides (see later Appendix sections for more details).

To finish preparing the training dataset, we excluded sequences that had more than fourteen nucleotides trimmed from either the V-gene or J-gene, as more extensive trimming is uncommon and could suggest an alternative trimming mechanism. We also focused only on non-productive sequences to avoid the potential confounding effects of selection in recombination statistics. After applying these filtering criteria, the final training dataset consisted of 1,257,528 sequences. To validate our model, we used a separate dataset of 983,514 productive sequences.

TCRα testing dataset

We downloaded TCRα repertoire sequence data from peripheral blood samples of 10 healthy individuals (aged 3-14 years) from the Adaptive Biotechnologies immuneACCESS database using the link in the original publication [119]. This cohort differs from the training cohort in demographics and sampling location. All individuals were heterozygous for HLA-DR3/DR4 and had siblings diagnosed with Type-1 Diabetes, but they showed no clinical symptoms of diabetes at the time of sampling or in the subsequent years. We applied the same IGoR-based annotation and filtering procedures to this testing dataset as used for the training dataset. For model validation, we separately analyzed 98,244 non-productive and 141,676 productive sequences.

TCRγ testing dataset

Annotated TCRγ repertoire sequence data for 23 healthy bone marrow donor subjects was downloaded from the Adaptive Biotechnologies immuneACCESS database [102]. We applied the same filtering procedures to this testing dataset as used for the training dataset. For model validation, we separately analyzed approximately 44,673 non-productive and 20,681 productive sequences.

F.2 Identifying the set of possible sequence trimming and ligation annotations

We aim to identify all feasible combinations of trimming and ligation values (delV, delJ, and MH) that could account for the observed sequence X. Recall that V and J represent the V-gene and J-gene, which each define a V-gene and J-gene sequence, respectively. For ease of notation, these sequences are both oriented in the 3'-to-5' direction and are represented as ordered lists of nucleotides. Although the top strand of the V-gene typically follows a 5'-to-3' orientation, we reverse it here for notational convenience.

Recall that each sequence receives an initial annotation inferred by IGoR, consisting of a V-gene and J-gene assignment, trimming scenario, and N-insertion amount. While we use the gene and N-insertion assignments directly, we do not use the IGoR-inferred trimming scenario as-is. This is because IGoR does not explicitly account for microhomology and assigns shared nucleotides to only one gene segment. Instead, we adapt the trimming annotations to incorporate possible germline-encoded microhomology, generating a set of possible trimming and ligation scenarios for each sequence, including those involving microhomologous nucleotides.

To begin, let delV_i and delJ_i represent specific V- and J-gene trimming amounts inferred by IGoR in the initial annotation. Since IGoR assumes MH = 0, the initial set of possible annotations, A_X, includes only $(\text{delV} = \text{delV}_i, \text{delJ} = \text{delJ}_i, \text{MH} = 0)$. To expand A_X to

include annotations with microhomology (MH > 0), we introduce a function $h(x, y)$, which quantifies the number of contiguous complementary nucleotides between two overlapping, equal-length sequence regions x and y. This function is defined as:

$$h(x,y) = \sum_{i=0}^{\text{len}(x)} \begin{cases} 1 & \text{if } x(j) \text{ is complementary to } y(j) \text{ for all } j \in \{0,\ldots,i\} \\ 0 & \text{otherwise.} \end{cases} \tag{F.1}$$

We apply this function to assess complementarity in overlapping regions between a V-gene sequence (V) and a J-gene sequence (J), aligning the sequences without gaps at the IGoR-inferred trimming sites delV_i and delJ_i (Figure E.12). The overlapping regions are:

1. The *V-gene:trimmed-J* region, where $\mathsf{seq}_{\texttt{trimmed}}(\text{J}, \text{delJ}_i)$ represents the trimmed J-gene sequence oriented 5'-to-3', and $\mathsf{seq}_{\texttt{overlap}}(\text{V}, \text{delV}_i, \text{delJ}_i)$ represents the overlapping V-gene sequence oriented 3'-to-5'.

2. The *J-gene:trimmed-V* region, with $\mathsf{seq}_{\texttt{trimmed}}(\text{V}, \text{delV}_i)$ and $\mathsf{seq}_{\texttt{overlap}}(\text{J}, \text{delJ}_i, \text{delV}_i)$ representing the trimmed V-gene and overlapping J-gene sequences oriented 5'-to-3' and 3'-to-5', respectively.

We define k_J and k_V as the counts of contiguous complementary nucleotides in these regions:

$$k_J = h\big(\mathsf{seq}_{\texttt{overlap}}(\text{V}, \text{delV}_i, \text{delJ}_i), \mathsf{seq}_{\texttt{trimmed}}(\text{J}, \text{delJ}_i)\big)$$

and

$$k_V = h\big(\mathsf{seq}_{\texttt{overlap}}(\text{J}, \text{delJ}_i, \text{delV}_i), \mathsf{seq}_{\texttt{trimmed}}(\text{V}, \text{delV}_i)\big).$$

Given these values, a sequence X with an initial IGoR-inferred trimming scenario ($\text{delV} = \text{delV}_i, \text{delJ} = \text{delJ}_i, \text{MH} = 0$) can also be annotated with microhomologous nucleotide counts MH ranging from 0 to $k_V + k_J$. For each of these values of MH, the corresponding trimming

amounts (delV and delJ) are adjusted accordingly to ensure the same observed sequence is generated. This expands the set of possible annotations A_X to:

$$A_X = \{(\text{delV}_i, \text{delJ}_i, 0)\} \cup \{(\text{delV}_n, \text{delJ}_n, m) \mid \text{conditions}\}.$$

The conditions are:

1. For each possible annotation, delV_n and delJ_n (realizations of delV and delJ) are within the range of initial IGoR-inferred trimming values adjusted by the contiguous nucleotide count: $\text{delV}_i - k_V \leq \text{delV}_n \leq \text{delV}_i$ and $\text{delJ}_i - k_J \leq \text{delJ}_n \leq \text{delJ}_i$.

2. The microhomologous nucleotide count m (a realization of MH) ranges from 0 to the sum of contiguous complementary nucleotides: $0 \leq m \leq k_V + k_J$.

3. The sum of the trimming amounts delV_n and delJ_n and the microhomologous nucleotide count m equals the sum of the initial IGoR-inferred trimming amounts: $\text{delV}_n + \text{delJ}_n + m = \text{delV}_i + \text{delJ}_i$

This process can be repeated to identify the sets of all possible sequence annotations for each sequence sampled from a TCRα repertoire.

F.3 Defining a model weight function

We aim to model the influence of various sequence-level parameters, including microhomology-related parameters, on joint trimming and ligation scenario probabilities, $P(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0)$, where delVJ represents the trimming scenario (defined by delV and delJ), MH represents the number of microhomologous nucleotides used in ligation, VJ represents the gene pair, Q represents the productivity of the sequences, and $\text{I} = 0$ represents zero N-insertions. For our modeling purposes, we assume the following about V(D)J recombination biology:

1. The DNA hairpin of each joining gene is nicked open by a single-stranded break [26, 27, 29–31].

2. This hairpin nick occurs at the +2 position, creating a 4-nucleotide-long 3'-single-stranded overhang, with the two 3'-most nucleotides being P-nucleotides [29, 31].

3. If any part of the original gene sequence is deleted, all P-nucleotides will also be deleted [26, 47].

These assumptions allow us to determine the germline nucleotide sequence on both sides of each trimming site and define sequence-level model features. We assume that observations can be drawn from a model where these features vary across trimming and/or ligation scenarios for a given gene pair. Using these assumptions, we previously demonstrated that local nucleotide identity surrounding trimming sites (the "trimming motif") and the counts of GC or AT nucleotides beyond these motifs (the "5' base-count" and "3' base-count") are highly predictive of trimming probabilities for single gene sequences [3]. Building on this foundation, we aim to integrate these established parameters with newly developed microhomology-related parameters to assess the combined effects on the processes of trimming and ligation.

For our model, we define two sets of parameters, one trimming-related and one ligation-related, to model the probabilities of trimming and ligation scenarios. We model trimming scenario probabilities using established trimming motif (given by $\boldsymbol{\beta}_V^{\texttt{motif}}$ and $\boldsymbol{\beta}_J^{\texttt{motif}}$) and 5' and 3' base count (given by $\boldsymbol{\beta}_V^{\texttt{AT}}$, $\boldsymbol{\beta}_J^{\texttt{AT}}$, $\boldsymbol{\beta}_V^{\texttt{GC}}$, and $\boldsymbol{\beta}_J^{\texttt{GC}}$) parameters for each gene, along with a new parameter related to microhomology (given by $\beta^{\texttt{trimMH}}$). This new parameter measures the trimming-related effect of the average number of microhomologous nucleotides across all possible ligation scenarios for a given trimming scenario. Additionally, we model ligation scenario probabilities using another new parameter related to microhomology (given by $\beta^{\texttt{ligMH}}$), which measures the ligation-related effect of the number of microhomologous nucleotides within the ligation scenario. Using these model parameters, we define weight

functions for the trimming choice $f_{\texttt{trim}}$ and the ligation choice $f_{\texttt{lig}}$ such that our model of $P(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0)$ will be a normalized version of these weights.

The trimming-related weight function $f_{\texttt{trim}}$ aggregates the desired parameter-specific weight functions (defined in Table E.2) as follows:

$$\begin{aligned} &f_{\texttt{trim}}(\text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{trim}}) \\ &\qquad \coloneqq f_{\texttt{motif}}(\text{delV}, \text{V}; \boldsymbol{\beta}_V^{\texttt{motif}}) + f_{\texttt{motif}}(\text{delJ}, \text{J}; \boldsymbol{\beta}_J^{\texttt{motif}}) \\ &\qquad + f_{\texttt{count}}(\text{delV}, \text{V}; \boldsymbol{\beta}_V^{\texttt{AT}}, \boldsymbol{\beta}_V^{\texttt{GC}}) + f_{\texttt{count}}(\text{delJ}, \text{J}; \boldsymbol{\beta}_J^{\texttt{AT}}, \boldsymbol{\beta}_J^{\texttt{GC}}) \\ &\qquad + f_{\texttt{trimMH}}(\text{delVJ}, \text{VJ}; \beta^{\texttt{trimMH}}). \end{aligned} \tag{F.2}$$

For notational convenience, $\boldsymbol{\beta}_{\texttt{trim}}$ represents the set of all trimming-related regression parameters, $\text{VJ} = (\text{V}, \text{J})$ represents a gene pair, $\text{delVJ} = (\text{delV}, \text{delJ})$ represents a trimming scenario, and MH represents the number of microhomologous nucleotides within the ligation scenario.

Additionally, we define a ligation-related weight function $f_{\texttt{lig}}$ that consists of the relevant parameter-specific weight function (defined in Table E.2) as follows:

$$f_{\texttt{lig}}(\text{delVJ}, \text{MH}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}}) \coloneqq f_{\texttt{ligMH}}(\text{MH}; \beta^{\texttt{ligMH}}) \tag{F.3}$$

where $\boldsymbol{\beta}_{\texttt{lig}}$ represents the set of all ligation-related regression parameters, which for our purposes includes only $\beta^{\texttt{ligMH}}$.

We further define each of these trimming-related and ligation-related regression parameters, along with their corresponding weight functions, within the following sections:

Extended parameter description

Defining "trimming motif" parameters: As in our previous work [3], we define trimming motif parameters to include one nucleotide position 5' of the trimming site and two nucleotide positions 3' of the trimming site. We describe this definition for a V-gene sequence and V-gene trimming amount, but it applies similarly to a J-gene sequence and J-gene trimming amount.

Recall that V represents the V-gene which defines a V-gene sequence. For ease of notation, we orient this sequence in the 3'-to-5' direction and represent it as an ordered list of nucleotides. Recall that delV is a random variable representing a V-gene trimming amount. Let $\mathrm{V}(\mathrm{delV}+2-j)$ represent the nucleotide identity at the trimming motif position $j \in \{0, \ldots, 2\}$ where positions $j \leq 0$ represent motif positions 5' of the trimming site and positions $j > 0$ represent motif positions 3' of the trimming site. The trimming motif sequence (oriented 5'-to-3') is given by the ordered list:

$$\big(\mathrm{V}(\mathrm{delV}+2-j)\big)_{j=0}^{2}. \tag{F.4}$$

Depending on delV, this trimming motif may or may not include P-nucleotides. For $\mathrm{delV} \geq 2$, the two 3' trimming motif nucleotides will include the two deleted gene sequence nucleotides 3' of the trimming site (and no P-nucleotides). Since we are assuming that the initial hairpin nick occurs at the $+2$ position, there will be two P-nucleotides present in the 5'-to-3' gene sequence. For $0 \leq \mathrm{delV} < 2$, P-nucleotides will be included in the trimming motif sequence. Likewise, as a result of the $+2$ hairpin nick position assumption, TCRs that have $\mathrm{delV} < 0$ will not have a full-length nucleotide trimming motif. For these "off-the-end" motif cases, we assign zero influence to the missing nucleotides during model fitting.

Let $\beta_{jk}^{\texttt{motif}}$ be a (log) position-weight-matrix parameter for trimming motif position $j \in \{0, \ldots, 2\}$ and nucleotide $k \in \{A, T, C, G\}$. The set of all such parameters for the V-gene is

denoted by $\boldsymbol{\beta}_V^{\texttt{motif}}$. We can define an un-normalized position-weight-matrix weight:

$$f_{\texttt{motif}}(\text{delV}, \text{V}; \boldsymbol{\beta}_V^{\texttt{motif}}) := \sum_{j=0}^{2} \beta_{j\text{V}(\text{delV}+2-j)}^{\texttt{motif}} \tag{F.5}$$

that will serve as a *motif*-specific weight function in subsequent modeling. As described above, since we are considering "off-the-end" motif cases, $\text{V}(\text{delV} + 2 - j)$ will represent the nucleotide identity at sequence position j where positions $j \leq 0$ represent sequence positions 5' of the trimming site and positions $j > 0$ represent sequence positions 3' of the trimming site.

Defining "base count" parameters: As in our previous work [3], we will also define parameters for the counts of GC and AT nucleotides on either side of each trimming site. We describe this definition for a V-gene sequence and V-gene trimming amount, but it applies similarly to a J-gene sequence and J-gene trimming amount. For an arbitrary sequence x, we can count the number of AT and GC nucleotides within the sequence as

$$C^{\texttt{AT}}(x) = C^{\texttt{A}}(x) + C^{\texttt{T}}(x) \tag{F.6}$$

and

$$C^{\texttt{GC}}(x) = C^{\texttt{G}}(x) + C^{\texttt{C}}(x), \tag{F.7}$$

respectively.

Since the count of AT or GC nucleotides within the sequences 5' and 3' of the trimming site may influence the probability of trimming differently, we calculate the counts separately and exclude nucleotides already included in the *motif* parameterization. As above, recall that delV represents a V-gene trimming amount and V represents a V-gene which defines a V-gene sequence. For ease of notation, we orient this sequence in the 3'-to-5' direction and

represent it as an ordered list of nucleotides. Let $\mathrm{V}(\mathrm{delV}+2-j)$ represent the nucleotide identity at the trimming motif position $j \in \{0,\ldots,2\}$ where positions $j \leq 0$ represent motif positions 5' of the trimming site and positions $j > 0$ represent motif positions 3' of the trimming site. As in our previous work, we include the ten nucleotides 5' of the motif in the 5' nucleotide counts. Since we include one nucleotide 5' of the trimming site in the "trimming motif" parameters (as described in the previous section), the nucleotide sequence 5' of the trimming site, beyond the "trimming motif", is given by the ordered list

$$\mathsf{seq}_5(\mathrm{delV},\mathrm{V}) = \big(\mathrm{V}(\mathrm{delV}+2-j)\big)_{j=-11}^{-1}. \tag{F.8}$$

To count the number of AT and GC nucleotides in the sequence 3' of the trimming site, we include all nucleotides located 3' of the trimming site beyond the "trimming motif." Since we are interested in using GC nucleotide content as a proxy for sequence-breathing, which is relevant only for paired nucleotides, we exclude nucleotides within the 3' single-stranded overhang. Assuming the initial hairpin nick occurs at the $+2$ position, leading to a 4-nucleotide-long 3' single-stranded overhang, for $\mathrm{delV} > 2$, the nucleotide sequence 3' of the trimming site, beyond the "trimming motif" (which contains two nucleotide positions 3' of the trimming site), is given by the ordered list:

$$\mathsf{seq}_3(\mathrm{delV},\mathrm{V}) = \begin{cases} \big(\mathrm{V}(\mathrm{delV}+2-j)\big)_{j=3}^{(\mathrm{delV}-2)} & \text{if } (\mathrm{delV}-2) \geq 3 \\ () & \text{if } (\mathrm{delV}-2) < 3. \end{cases} \tag{F.9}$$

For $(\mathrm{delV}-2) < 3$, all nucleotides 3' of the trimming site are considered single-stranded, and thus no nucleotides will be included in the sequence used to calculate the AT and GC base-counts.

With these sequences 5' and 3' of the trimming site, we define $\beta_{5V}^{\mathtt{AT}}$, $\beta_{3V}^{\mathtt{AT}}$, $\beta_{5V}^{\mathtt{GC}}$, and $\beta_{3V}^{\mathtt{GC}}$ to be V-gene specific *base count* model parameters for 5' and 3' sequence base-counts of AT

and GC beyond the "trimming motif", respectively. The set of all such parameters for the V-gene are denoted by $\boldsymbol{\beta}_V^{\mathtt{AT}}$ and $\boldsymbol{\beta}_V^{\mathtt{GC}}$. With these parameters, we define a *base count* weight function:

$$\begin{aligned} f_{\mathtt{count}}(\text{delV}, \text{V}; \boldsymbol{\beta}_V^{\mathtt{AT}}, \boldsymbol{\beta}_V^{\mathtt{GC}}) := \beta_{5V}^{\mathtt{AT}} \cdot C^{\mathtt{AT}}(\mathsf{seq}_5(\text{delV}, \text{V})) + \beta_{3V}^{\mathtt{AT}} \cdot C^{\mathtt{AT}}(\mathsf{seq}_3(\text{delV}, \text{V})) \\ + \beta_{5V}^{\mathtt{GC}} \cdot C^{\mathtt{GC}}(\mathsf{seq}_5(\text{delV}, \text{V})) + \beta_{3V}^{\mathtt{GC}} \cdot C^{\mathtt{GC}}(\mathsf{seq}_3(\text{delV}, \text{V})). \end{aligned} \tag{F.10}$$

using the functions $C^{\mathtt{AT}}$ and $C^{\mathtt{GC}}$ as defined in (F.6) and (F.7), respectively. As defined, these GC and AT base-counts for the 3' sequence are dependent on sequence length and provide a parameterization of both GC nucleotide content and length.

Defining "microhomology" parameters for trimming scenario choice: We can parameterize the average number of microhomologous nucleotides across possible ligation scenario choices for a given trimming scenario and define $\beta^{\mathtt{trimMH}}$ to be an *microhomology* model parameter specific to trimming choice. We define a function g that returns this average value as follows:

$$g(\text{delVJ}, \text{VJ}) := \frac{\sum_{\text{MH}' \in \mathcal{M}_{\text{VJ,delVJ}}} \text{MH}'}{|\mathcal{M}_{\text{VJ,delVJ}}|}$$

such that $\mathcal{M}_{\text{VJ,delVJ}}$ is the set of all possible ligation scenarios for the chosen trimming scenario delVJ and gene pair VJ. With this function, we can define a *microhomology* weight function

$$f_{\mathtt{trimMH}}(\text{delVJ}, \text{VJ}; \beta^{\mathtt{trimMH}}) := \beta^{\mathtt{trimMH}} \cdot g(\text{delVJ}, \text{VJ}). \tag{F.11}$$

Defining "microhomology" parameters for ligation scenario choice: We can directly use MH, which represents the number of microhomologous nucleotides in an observed ligation scenario, as a parameter. We then define $\beta^{\mathtt{ligMH}}$ as the *microhomology* model parameter for predicting the choice of ligation scenario. We use the term "observed" for this microhomology parameter because these particular microhomologous nucleotides directly participate

in the ligation process and are homologous in the final sequence. As such, we can define a *microhomology* weight function:

$$f_{\texttt{ligMH}}(\text{MH}; \beta^{\texttt{ligMH}}) \coloneqq \beta^{\texttt{ligMH}} \cdot \text{MH}. \tag{F.12}$$

F.4 Extended model formulation and training description

We aim to model the influence of various sequence-level parameters, including microhomology-related parameters, on joint trimming and ligation scenario probabilities, $P(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0)$, where delVJ represents the trimming scenario, MH represents the number of microhomologous nucleotides used in ligation, VJ represents the gene pair, Q represents the productivity of the sequences, and $\text{I} = 0$ represents zero N-insertions. Modeling this probability is complex because the true trimming and ligation annotation of each sampled sequence is a latent variable that will depend on the model parameters. As described earlier, we obtain the set of possible trimming and ligation scenario annotations (delVJ and MH), denoted as A_X, for a given sequence X by transforming the initial IGoR-inferred annotation. We assign probabilities (or weights) to each potential annotation, and since these probabilities depend on the model parameters, we use an expectation-maximization algorithm for parameter inference. Below, we provide a detailed description of these steps.

We employ a two-step conditional logit model to capture the decision-making involved in selecting trimming and ligation scenarios for V-J gene pairs. Our model describes a generative process in two steps:

1. We model the probability, $P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} = 0)$, of choosing a trimming scenario delVJ for a given V-J gene pair VJ, sequence productivity Q, and N-insertion amount $\text{I} = 0$. This probability is modeled by parameters specific to trimming scenarios.

2. We model the probability, $P(\text{MH} \mid \text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0)$, of choosing a ligation scenario

MH for a given trimming scenario delVJ, V-J gene pair VJ, sequence productivity Q, and N-insertion amount I = 0. This probability is modeled by parameters specific to each ligation scenario.

The joint probability of a trimming scenario delVJ and a ligation scenario MH for a given V-J gene pair VJ, sequence productivity Q, and N-insertion amount I = 0 can be factored as:

$$P(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0) = P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} = 0) \times P(\text{MH} \mid \text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0).$$

Figure 2 depicts the two-step structure of our model, illustrating the decision-making process for an example V-J gene pair.

To incorporate characteristics of each possible trimming and ligation scenario in our model, we define parameter-specific weight functions such that our model of $P(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0)$ will be a normalized version of these weights. In our previous work, we established that local nucleotide identities at trimming sites (the "trimming motif") and the counts of GC or AT nucleotides beyond these motifs (the "5' base-count" and "3' base-count") are strong predictors of trimming probabilities for single gene sequences [3]. Building on this foundation, we have integrated these established parameters with newly developed microhomology-related parameters to assess the combined effects on the processes of trimming and ligation. First, we define the trimming-related weight function $f_{\texttt{trim}}(\text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{trim}})$ parameterized by as set of trimming-related parameters $\boldsymbol{\beta}_{\texttt{trim}}$, which includes previously established trimming motif and base count parameters, and a new trimming-related microhomology parameter. This new parameter measures the effect of the average number of microhomologous nucleotides between two sequences, a value that varies depending on the chosen trimming scenario. Similarly, we define the ligation-related weight function $f_{\texttt{lig}}(\text{delVJ}, \text{MH}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})$ parameterized by a new ligation-related microhomology param-

eter $\boldsymbol{\beta}_{\texttt{lig}}$ which measures the effect of the number of microhomologous nucleotides between two sequences, a value that varies depending on the chosen trimming and ligation scenario. These parameters and weight functions are summarized in Table E.2 and defined in detail in previous Appendix sections.

With these weight functions, our model estimates the joint probability of a trimming and ligation scenario (given by delVJ and MH) for a given V-J gene pair VJ, sequence productivity Q, and N-insertion amount $I = 0$, combining influences of regression parameters $\boldsymbol{\beta}_{\texttt{trim}}$ and $\boldsymbol{\beta}_{\texttt{lig}}$:

$$P(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, I = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}}) := \\ P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}}) \times P(\text{MH} \mid \text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{lig}}). \tag{F.13}$$

To model the trimming scenario probability $P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}})$, we expand conditional probability, giving:

$$P(\text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}}) = P(\text{Q}, \text{I} = 0 \mid \text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}}) \times P(\text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{trim}}).$$

This probability $P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} = 0)$ is parameterized by both trimming- and ligation-related parameters ($\boldsymbol{\beta}_{\texttt{trim}}$ and $\boldsymbol{\beta}_{\texttt{lig}}$) because the model is conditioned on sequence productivity (Q), which is jointly determined by trimming and ligation. This dependency ensures that trimming probabilities properly account for how productivity constraints prune the space of possible ligation scenarios associated with each trimming scenario, correcting for any biases introduced by this non-uniform pruning. As such, we model $P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} =$

$0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}})$ as:

$$
\begin{aligned}
&P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}}) \\
&\quad = \frac{P(\text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}})}{\sum_{\text{delVJ}' \in \mathcal{D}} P(\text{delVJ}', \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{trim}}, \boldsymbol{\beta}_{\texttt{lig}})} \\
&\quad = \frac{P(\text{Q}, \text{I} = 0 \mid \text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}}) \cdot P(\text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{trim}})}{\sum_{\text{delVJ}' \in \mathcal{D}} P(\text{Q}, \text{I} = 0 \mid \text{delVJ}', \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}}) \cdot P(\text{delVJ}', \text{VJ}; \boldsymbol{\beta}_{\texttt{trim}})} \\
&\quad := \frac{P(\text{Q}, \text{I} = 0 \mid \text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}}) \cdot \exp\left(f_{\texttt{trim}}(\text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{trim}})\right)}{\sum_{\text{delVJ}' \in \mathcal{D}} P(\text{Q}, \text{I} = 0 \mid \text{delVJ}', \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}}) \cdot \exp\left(f_{\texttt{trim}}(\text{delVJ}', \text{VJ}; \boldsymbol{\beta}_{\texttt{trim}})\right)}
\end{aligned}
\tag{F.14}
$$

where $f_{\texttt{trim}}$ is the trimming-related weight defined in (F.2) and $\mathcal{D}$ is the set of all possible trimming scenarios for the specified sequence productivity Q and N-insertion amount $\text{I} = 0$. We model $P(\text{Q}, \text{I} = 0 \mid \text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})$ as:

$$
\begin{aligned}
P(\text{Q}, \text{I} = 0 | \text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}}) &= \frac{P(\text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{lig}})}{P(\text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})} \\
&= \frac{\sum_{\text{MH}_1 \in \mathcal{M}_1} P(\text{MH}_1, \text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{lig}})}{\sum_{\text{MH}_2 \in \mathcal{M}_2} P(\text{MH}_2, \text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})} \\
&= \frac{\sum_{\text{MH}_1 \in \mathcal{M}_1} P(\text{MH}_1, \text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})}{\sum_{\text{MH}_2 \in \mathcal{M}_2} P(\text{MH}_2, \text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})} \\
&:= \frac{\sum_{\text{MH}_1 \in \mathcal{M}_1} \exp\left(f_{\texttt{lig}}(\text{delVJ}, \text{MH}_1, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})\right)}{\sum_{\text{MH}_2 \in \mathcal{M}_2} \exp\left(f_{\texttt{lig}}(\text{delVJ}, \text{MH}_2, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})\right)}
\end{aligned}
\tag{F.15}
$$

given that $P(\text{MH}, \text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{lig}}) = P(\text{MH}, \text{delVJ}, \text{VJ}; \boldsymbol{\beta}_{\texttt{lig}})$. Here, $f_{\texttt{lig}}$ is the ligation-related weight defined in (F.12), $\mathcal{M}_1$ is the set of all possible ligation scenarios for the chosen trimming scenario delVJ, sequence productivity Q, and N-insertion amount $\text{I} = 0$ and $\mathcal{M}_2$ is the set of all possible ligation scenarios for the chosen trimming scenario delVJ.

Similarly, we model the ligation scenario probability $P(\text{MH} \mid \text{delVJ}, \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}_{\texttt{lig}})$

as:

$$
\begin{aligned}
P(\mathrm{MH} \mid \mathrm{delVJ}, \mathrm{VJ}, \mathrm{Q}, \mathrm{I}=0; \boldsymbol{\beta}_{\mathtt{lig}}) &= \frac{P(\mathrm{MH}, \mathrm{delVJ}, \mathrm{VJ}, \mathrm{Q}, \mathrm{I}=0; \boldsymbol{\beta}_{\mathtt{lig}})}{\sum_{\mathrm{MH}_1 \in \mathcal{M}_1} P(\mathrm{MH}_1, \mathrm{delVJ}, \mathrm{VJ}, \mathrm{Q}, \mathrm{I}=0; \boldsymbol{\beta}_{\mathtt{lig}})} \\
&= \frac{P(\mathrm{MH}, \mathrm{delVJ}, \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{lig}})}{\sum_{\mathrm{MH}_1 \in \mathcal{M}_1} P(\mathrm{MH}_1, \mathrm{delVJ}, \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{lig}})} \\
&:= \frac{\exp\left(f_{\mathtt{lig}}(\mathrm{delVJ}, \mathrm{MH}, \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{lig}})\right)}{\sum_{\mathrm{MH}_1 \in \mathcal{M}_1} \exp\left(f_{\mathtt{lig}}(\mathrm{delVJ}, \mathrm{MH}_1, \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{lig}})\right)}.
\end{aligned} \tag{F.16}
$$

Combining these, our model becomes

$$
\begin{aligned}
&P(\mathrm{delVJ}, \mathrm{MH} \mid \mathrm{VJ}, \mathrm{Q}, \mathrm{I}=0; \boldsymbol{\beta}_{\mathtt{lig}}, \boldsymbol{\beta}_{\mathtt{trim}}) \\
&\quad := P(\mathrm{delVJ} \mid \mathrm{VJ}, \mathrm{Q}, \mathrm{I}=0; \boldsymbol{\beta}_{\mathtt{trim}}, \boldsymbol{\beta}_{\mathtt{lig}}) \times P(\mathrm{MH} \mid \mathrm{delVJ}, \mathrm{VJ}, \mathrm{Q}, \mathrm{I}=0; \boldsymbol{\beta}_{\mathtt{lig}}) \\
&\quad := \frac{P(\mathrm{Q}, \mathrm{I}=0 \mid \mathrm{delVJ}, \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{lig}}) \cdot \exp\left(f_{\mathtt{trim}}(\mathrm{delVJ}, \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{trim}})\right)}{\sum_{\mathrm{delVJ}' \in \mathcal{D}} P(\mathrm{Q}, \mathrm{I}=0 \mid \mathrm{delVJ}', \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{lig}}) \cdot \exp\left(f_{\mathtt{trim}}(\mathrm{delVJ}', \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{trim}})\right)} \\
&\qquad\qquad \times \frac{\exp\left(f_{\mathtt{lig}}(\mathrm{delVJ}, \mathrm{MH}, \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{lig}})\right)}{\sum_{\mathrm{MH}_1 \in \mathcal{M}_1} \exp\left(f_{\mathtt{lig}}(\mathrm{delVJ}, \mathrm{MH}_1, \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{lig}})\right)}.
\end{aligned} \tag{F.17}
$$

where $P(\mathrm{Q}, \mathrm{I}=0 \mid \mathrm{delVJ}, \mathrm{VJ}; \boldsymbol{\beta}_{\mathtt{lig}})$ is defined in (F.15).

Recall that our data consists of sequences and we are considering sets of all possible sequence annotations A_X for each sampled sequence X. Each annotation includes a trimming scenario delVJ and ligation scenario MH. To infer the parameters of our model $P(\mathrm{delVJ}, \mathrm{MH} \mid \mathrm{VJ}, \mathrm{Q}, \mathrm{I}=0; \boldsymbol{\beta}_{\mathtt{lig}}, \boldsymbol{\beta}_{\mathtt{trim}})$, defined in F.17, while marginalizing over all possible sequence annotations for each sequence, we employ an expectation-maximization (EM) approach. This iterative algorithm proceeds as follows: starting with initial model parameters $\boldsymbol{\beta}_{\mathtt{lig}}$ and $\boldsymbol{\beta}_{\mathtt{trim}}$, we aim to update to improved parameters $\boldsymbol{\beta}'_{\mathtt{lig}}$ and $\boldsymbol{\beta}'_{\mathtt{trim}}$. We define the normalized conditional probability of a specific sequence annotation ($\mathrm{delVJ} = \mathrm{delVJ}_1, \mathrm{MH} =$

$\mathrm{MH}_1) \in A_X$ given a sequence X with gene pair $\mathrm{VJ} = \mathrm{VJ}_1$ as:

$$P_{\text{annot}}(\text{delVJ} = \text{delVJ}_1, \text{MH} = \text{MH}_1 \mid \text{VJ} = \text{VJ}_1, \text{Q} = \text{Q}_1, \text{I} = 0; X, \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}}) = \frac{P(\text{delVJ}_1, \text{MH}_1 \mid \text{VJ}_1, \text{Q}_1, I = 0; \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}})}{\sum_{(\text{delVJ}_n, \text{MH}_n) \in A_X} P(\text{delVJ}_n, \text{MH}_n \mid \text{VJ}_1, \text{Q}_1, I = 0; \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}})}. \tag{F.18}$$

Here, $P(\text{delVJ}_1, \text{MH}_1 \mid \text{VJ}_1, \text{Q}_1, I = 0; \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}})$ is computed according to (F.17). With this, we then define the expected log-likelihood of new parameter estimates $\boldsymbol{\beta}'_{\texttt{lig}}$ and $\boldsymbol{\beta}'_{\texttt{trim}}$ given the current estimates $\boldsymbol{\beta}_{\texttt{lig}}$ and $\boldsymbol{\beta}_{\texttt{trim}}$ for a single sampled sequence X as:

$$\begin{aligned}\ell(\boldsymbol{\beta}'_{\texttt{lig}}, \boldsymbol{\beta}'_{\texttt{trim}} \mid \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}}; X, \text{Q}, \text{I} = 0) \\ = \sum_{(\text{delVJ}_n, \text{MH}_n) \in A_X} & P_{\text{annot}}(\text{delVJ} = \text{delVJ}_n, \text{MH} = \text{MH}_n \mid \text{VJ} = \text{VJ}_n, \text{Q}, \text{I} = 0; X, \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}}) \\ & \times \log P(\text{delVJ} = \text{delVJ}_n, \text{MH} = \text{MH}_n \mid \text{VJ} = \text{VJ}_n, \text{Q}, \text{I} = 0; \boldsymbol{\beta}'_{\texttt{lig}}, \boldsymbol{\beta}'_{\texttt{trim}})\end{aligned} \tag{F.19}$$

where $P_{\text{annot}}(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0; X, \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}})$ and $P(\text{delVJ}, \text{MH} \mid \text{VJ}, \text{Q}, \text{I} = 0; \boldsymbol{\beta}'_{\texttt{lig}}, \boldsymbol{\beta}'_{\texttt{trim}})$ are defined in (F.18) and (F.17), respectively. Similarly, we define the log-likelihood function for a random sample of observed sequences $\mathcal{X}$ as:

$$\mathcal{L}(\boldsymbol{\beta}'_{\texttt{lig}}, \boldsymbol{\beta}'_{\texttt{trim}} \mid \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}}; \mathcal{X}, \text{Q}, \text{I} = 0) = \sum_{X \in \mathcal{X}} C(X) \times \ell(\boldsymbol{\beta}'_{\texttt{lig}}, \boldsymbol{\beta}'_{\texttt{trim}} \mid \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}}; X, \text{Q}, \text{I} = 0) \tag{F.20}$$

where $C(X)$ represents the observed count of a specific sequence $X \in \mathcal{X}$ in the sampled data and $\ell(\boldsymbol{\beta}'_{\texttt{lig}}, \boldsymbol{\beta}'_{\texttt{trim}} \mid \boldsymbol{\beta}_{\texttt{lig}}, \boldsymbol{\beta}_{\texttt{trim}}; X, \text{Q}, \text{I} = 0)$ is defined as in (F.19). The calculation of this expectation $\mathcal{L}$ constitutes the E-step of our EM procedure. Subsequently, in the minimization step (M-step), we update the model parameters by minimizing the negative log-likelihood of the observed data obtained in the E-step. This minimization step is performed using gradient descent with the `JAX` and `JAXopt` packages in Python [121, 122]. The algorithm iterates between the E and M steps until changes in the negative log-likelihood between

successive iterations fall below a predefined threshold, indicating convergence.

F.5 Evaluating model using simulated data

To ensure our model returned expected outputs, we designed a data simulator capable of generating data under specific microhomology regimes. The simulator first samples a V-gene and J-gene according to IGoR-derived gene usage probabilities. Next, we establish probabilities for each trimming scenario using outputs from a version of our model that excludes microhomology terms, incorporating only trimming motif and base count terms. We then adjust these trimming probabilities using a tunable parameter to simulate the effect of microhomology, and the simulator samples a trimming scenario based on these adjusted probabilities. Finally, the simulator samples a ligation scenario uniformly, unless adjusted for microhomology effects by another tunable parameter. These two tunable parameters allow us to control the influence of microhomology on trimming and ligation choices. This process generates an observed simulated sequence, and by repeating it, we obtain a large set of simulated sequences to train and evaluate our model. We ran the simulator in four modes:

1. **No microhomology effect:** Both tunable microhomology parameters set to zero; microhomology does not influence trimming or ligation choices.

2. **Microhomology affects both trimming and ligation:** Both tunable microhomology parameters set to nonzero, positive values; microhomology increases probabilities for both trimming and ligation choices.

3. **Microhomology affects trimming, but not ligation:** Trimming-related parameter set to a nonzero, positive value and ligation-related parameter set to zero; microhomology increases trimming choice probabilities but does not affect ligation probabilities.

4. **Microhomology affects ligation, but not trimming:** Ligation-related parameter

set to a nonzero, positive value and trimming-related parameter set to zero; microhomology increases ligation choice probabilities but does not affect trimming probabilities.

Using these simulated datasets, we trained our model to ensure that the expected inferred parameters were obtained. We also adjusted the strength of the microhomology-related effects using these tunable parameters to ensure our model could capture these signals.

F.6 Exploring the relationship between microhomology and trimming probabilities, independent of ligation

To quantify the effect of microhomology on trimming scenario probabilities independently of ligation, we restrict our training dataset to non-productive sequences *containing* N-insertions, as their presence suggests that ligation exclusively involving germline microhomology did not occur. We aim to determine the influence of various sequence-level parameters on $P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} > 0)$, where delVJ represents the trimming scenario, VJ represents the gene pair, Q represents the sequence productivity, and $\text{I} > 0$ represents nonzero N-insertions.

We previously demonstrated that local nucleotide identity surrounding trimming sites (the "trimming motif") and the counts of GC or AT nucleotides beyond these motifs (the "5' base-count" and "3' base-count") are highly predictive of trimming probabilities for single gene sequences [3]. Here, we model the probabilities of trimming scenarios for gene pairs using these established parameters along with a new parameter related to "intermediate microhomology." This new parameter measures the importance of the average number of microhomologous nucleotides between two trimmed sequences, which, notably, are not homologous in the final rearranged sequence, see following section for definition. A summary of these model parameters and their corresponding weights for an arbitrary gene pair $\text{VJ} = (\text{V}, \text{J})$ and trimming scenario $\text{delVJ} = (\text{delV}, \text{delJ})$ is given in Table E.2.

Using these model features, we define a weight function f such that our model of $P(\text{delVJ}_i \mid \text{VJ}, \text{Q}, \text{I} > 0)$ will be a normalized version of this weight. We parameterize f

using $\boldsymbol{\beta}$, the set of all model parameters, as follows:

$$\begin{aligned} f(\text{delVJ}, \text{VJ}; \boldsymbol{\beta}) &\coloneqq f(\text{delVJ}, \text{VJ}; \boldsymbol{\beta}_V^{\texttt{motif}}, \boldsymbol{\beta}_J^{\texttt{motif}}, \boldsymbol{\beta}_V^{\texttt{AT}}, \boldsymbol{\beta}_V^{\texttt{GC}}, \boldsymbol{\beta}_J^{\texttt{AT}}, \boldsymbol{\beta}_J^{\texttt{GC}}, \beta^{\texttt{iMH}}) \\ &\coloneqq f_{\texttt{motif}}(\text{delV}, \text{V}; \boldsymbol{\beta}_V^{\texttt{motif}}) + f_{\texttt{motif}}(\text{delJ}, \text{J}; \boldsymbol{\beta}_J^{\texttt{motif}}) \\ &\quad + f_{\texttt{count}}(\text{delV}, \text{V}; \boldsymbol{\beta}_V^{\texttt{AT}}, \boldsymbol{\beta}_V^{\texttt{GC}}) + f_{\texttt{count}}(\text{delJ}, \text{J}; \boldsymbol{\beta}_J^{\texttt{AT}}, \boldsymbol{\beta}_J^{\texttt{GC}}) \\ &\quad + f_{\texttt{iMH}}(\text{delV}, \text{delJ}, \text{V}, \text{J}; \beta^{\texttt{iMH}}). \end{aligned} \tag{F.21}$$

Here, the parameter-specific weights $f_{\texttt{motif}}$, $f_{\texttt{count}}$, and $f_{\texttt{iMH}}$ are summarized in Table E.2 and in the following section.

With this weight formulation, we can fit a conditional logit model which posits

$$P(\text{delVJ} \mid \text{VJ}, \text{Q}, \text{I} > 0; \boldsymbol{\beta}) \coloneqq \frac{\exp\big(f(\text{delVJ}, \text{VJ}; \boldsymbol{\beta})\big)}{\sum_{\text{delVJ}_n \in \mathcal{D}} \exp\big(f(\text{delVJ}_n, \text{VJ}; \boldsymbol{\beta})\big)}. \tag{F.22}$$

Here, VJ and delVJ are random variables representing the V-gene and J-gene pair and trimming scenario, respectively, and $\mathcal{D}$ is the set of all reasonable trimming scenarios.

The likelihood function $\ell(\boldsymbol{\beta})$ for a random sample of sequences is the likelihood of the model parameters $\boldsymbol{\beta}$ given a set of observed trimming scenarios for specific gene pairs. The log likelihood function is:

$$\begin{aligned} &\log \ell(\boldsymbol{\beta}) \\ &\quad = \sum_{\text{VJ}_n \in \mathcal{S}} \sum_{\text{delVJ}_n \in \mathcal{D}} C(\text{delVJ}_n, \text{VJ}_n, \text{Q}, \text{I} > 0) \cdot \log P(\text{delVJ}_n \mid \text{VJ}_n, \text{Q}, \text{I} > 0; \boldsymbol{\beta}) \end{aligned} \tag{F.23}$$

where $\mathcal{S}$ represents the set of all gene pairs and $\mathcal{D}$ represents the set of all reasonable IGoR-inferred trimming scenarios. Here, $C(\text{delVJ}_n, \text{VJ}_n, \text{Q}, \text{I} > 0)$ is the count of sequences with IGoR-inferred trimming scenario delVJ_n, gene pair VJ_n, nonzero N-insertions, and sequence productivity Q and $P(\text{delVJ}_n \mid \text{VJ}_n, \text{Q}, \text{I} > 0; \boldsymbol{\beta})$ is given by Equation (F.22).

We include an additional regularization term for the intermediate-microhomology-specific

parameters to help prevent over-fitting during model training. As such, we define a log loss function as

$$\mathcal{L}(\boldsymbol{\beta}) = -\log \ell(\boldsymbol{\beta}) + \lambda \cdot (\beta^{\texttt{iMH}})^2 \tag{F.24}$$

where $\log \ell(\boldsymbol{\beta})$ is given by (F.23) and λ is a L2 regularization hyperparameter. We minimize this log loss function using gradient descent with the `JAX` and `JAXopt` packages in Python [121, 122]. We use a grid search to optimize the L2 regularization hyperparameter, λ. Notably, training this model without regularization (i.e. $\lambda = 0$) yields the same results as using the `mclogit` package in R, another implementation of conditional logistic regression that does not allow for regularization.

Defining "intermediate microhomology" parameters

In addition to the previously defined parameters, we define new microhomology-related parameters to model possible intermediate effects of microhomology on the observed trimming scenario, specifically when exploring the relationship between microhomology and trimming probabilities independent of ligation (as described above). We use the term "intermediate" because these nucleotides, while not directly participating in the final ligation, may temporarily influence intermediate steps such as trimming. Let a be a non-negative integer value that represents the number of nucleotides 5' of each trimming site which are allowed to overlap between the two sequences when orienting the top strand of the V-gene sequence 5'-to-3' and the bottom strand of the J-gene sequence 3'-to-5' (e.g. as highlighted in yellow in Figure E.13). Given a value of a, random variables V and J representing a V-gene and J-gene which each define a V-gene and J-gene sequence (both oriented 3'-to-5' as ordered lists), and random variables delV_i and delJ_i representing V-gene and J-gene trimming amounts (inferred directly from IGoR), the sub-sequences corresponding to this overlapping

region are defined by the following ordered lists

$$\mathsf{seq}_{\texttt{Vmh}}(\mathrm{V}, \mathrm{delV}_i, a) = \begin{cases} \big(\mathrm{V}(\mathrm{delV}_i + 2 - j)\big)_{j=(1-a)}^{0} & \text{if } a \geq 1 \\ () & \text{if } a = 0 \end{cases} \tag{F.25}$$

and

$$\mathsf{seq}_{\texttt{Jmh}}(\mathrm{J}, \mathrm{delJ}_i, a) = \begin{cases} \big(\mathrm{J}(\mathrm{delJ}_i + 2 - j)\big)_{j=0}^{(1-a)} & \text{if } a \geq 1 \\ () & \text{if } a = 0. \end{cases} \tag{F.26}$$

Here, $\mathrm{V}(\mathrm{delV}_i + 2 - j)$ and $\mathrm{J}(\mathrm{delJ}_i + 2 - j)$ represent the nucleotide identities at a sequence position j where positions $j \leq 0$ represent sequence positions 5' of the trimming sites and positions $j > 0$ represent sequence positions 3' of the trimming sites. The resulting sub-sequences, $\mathsf{seq}_{\texttt{Vmh}}(\mathrm{V}, \mathrm{delV}_i, a)$ and $\mathsf{seq}_{\texttt{Jmh}}(\mathrm{J}, \mathrm{delJ} + i, a)$, are oriented in the 5'-to-3' and 3'-to-5' directions, respectively, making them complementary. To quantify microhomology, we can define a function g which will count the number of complementary (i.e. microhomologous) nucleotides between two arbitrary overlapping, equal-length sequence regions, x and y, as

$$g(x, y) = \sum_{i=0}^{\mathrm{len}(x)} \begin{cases} 1 & \text{if } x(i) \text{ is complementary to } y(i) \\ 0 & \text{otherwise.} \end{cases} \tag{F.27}$$

It has been established that classical non-homologous end joining, which is the joining process used during V(D)J recombination, may involve up to four nucleotides of microhomology [109]. As such, we quantify the average number of non-contiguous microhomologous nucleotides across these overlapping interior sub-sequences corresponding to each $a \in \{1, 2, 3, 4\}$ as follows:

$$m(\mathrm{V}, \mathrm{J}, \mathrm{delV}_i, \mathrm{delJ}_i) := \frac{\sum_{a \in \{1,2,3,4\}} g\big(\mathsf{seq}_{\texttt{Vmh}}(\mathrm{V}, \mathrm{delV}_i, a), \mathsf{seq}_{\texttt{Jmh}}(\mathrm{J}, \mathrm{delJ}_i, a)\big)}{4} \tag{F.28}$$

With this average number of microhomologous nucleotides, we define an *intermediate microhomology* model parameter, $\beta^{\texttt{iMH}}$, and a corresponding weight function for a pair of IGoR-inferred trimming sites delV_i and delJ_i and genes V and J:

$$f_{\texttt{iMH}}(\text{delV}_i, \text{delJ}_i, \text{V}, \text{J}; \beta^{\texttt{iMH}}) \coloneqq \beta^{\texttt{iMH}} \cdot m(\text{V}, \text{J}, \text{delV}_i, \text{delJ}_i) \tag{F.29}$$

using the previously defined sequences and the function m as defined in (F.28).

BIBLIOGRAPHY

[1] Magdalena L Russell, Assya Trofimov, Philip Bradley, and Frederick A Matsen, 4th. Statistical analysis of repertoire data demonstrates the influence of microhomology in V(D)J recombination. *bioRxiv*, October 2024.

[2] Magdalena L Russell, Aisha Souquette, David M Levine, Stefan A Schattgen, E Kaitlynn Allen, Guillermina Kuan, Noah Simon, Angel Balmaseda, Aubree Gordon, Paul G Thomas, Frederick A Matsen, 4th, and Philip Bradley. Combining genotypes and T cell receptor distributions to infer genetic loci determining V(D)J recombination probabilities. *Elife*, 11, March 2022.

[3] Magdalena L Russell, Noah Simon, Philip Bradley, and Frederick A Matsen, 4th. Statistical inference reveals the role of length, GC content, and local sequence in V(D)J nucleotide trimming. *Elife*, 12, May 2023.

[4] Adrien Six, Maria Encarnita Mariotti-Ferrandiz, Wahiba Chaara, Susana Magadan, Hang-Phuong Pham, Marie-Paule Lefranc, Thierry Mora, Véronique Thomas-Vaslin, Aleksandra M Walczak, and Pierre Boudinot. The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Front. Immunol.*, 4:413, November 2013.

[5] Edus H Warren, Frederick A Matsen, 4th, and Jeffrey Chou. High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. *Blood*, 122(1):19–22, July 2013.

[6] Daniel J Woodsworth, Mauro Castellarin, and Robert A Holt. Sequence analysis of t-cell repertoires in health and disease. *Genome Med.*, 5(10):98, October 2013.

[7] George Georgiou, Gregory C Ippolito, John Beausang, Christian E Busse, Hedda Wardemann, and Stephen R Quake. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.*, 32(2):158–168, February 2014.

[8] Xavier Brochet, Marie-Paule Lefranc, and Véronique Giudicelli. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.*, 36(Web Server issue):W503–8, July 2008.

[9] Jian Ye, Ning Ma, Thomas L Madden, and James M Ostell. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.*, 41(Web Server issue): W34–40, July 2013.

[10] Dmitriy A Bolotin, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*, 12(5):380–381, May 2015.

[11] Quentin Marcou, Thierry Mora, and Aleksandra M Walczak. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.*, 9(1), December 2018.

[12] Li Song, David Cohen, Zhangyi Ouyang, Yang Cao, Xihao Hu, and X Shirley Liu. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods*, 18(6):627–630, June 2021.

[13] Victor Greiff, Enkelejda Miho, Ulrike Menzel, and Sai T Reddy. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.*, 36(11):738–749, November 2015.

[14] Grégoire Altan-Bonnet, Thierry Mora, and Aleksandra M Walczak. Quantitative immunology for physicists. *Phys. Rep.*, 849:1–83, March 2020.

[15] Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan, Jr. Statistical inference of the generation probability of t-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U. S. A.*, 109(40):16161–16166, October 2012.

[16] Yuval Elhanati, Anand Murugan, Curtis G Callan, Jr, Thierry Mora, and Aleksandra M Walczak. Quantifying selection in immune receptor repertoires. *Proc. Natl. Acad. Sci. U. S. A.*, 111(27):9875–9880, July 2014.

[17] Zachary Sethna, Giulio Isacchini, Thomas Dupic, Thierry Mora, Aleksandra M Walczak, and Yuval Elhanati. Population variability in the generation and selection of t-cell repertoires. *PLoS Comput. Biol.*, 16(12):e1008394, December 2020.

[18] Giulio Isacchini, Aleksandra M Walczak, Thierry Mora, and Armita Nourmohammad. Deep generative selection models of T and B cell receptor repertoires with soNNia. *Proc. Natl. Acad. Sci. U. S. A.*, 118(14), April 2021.

[19] Harlan S Robins, Santosh K Srivastava, Paulo V Campregher, Cameron J Turtle, Jessica Andriesen, Stanley R Riddell, Christopher S Carlson, and Edus H Warren. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci. Transl. Med.*, 2(47):47ra64, September 2010.

[20] Zachary Sethna, Yuval Elhanati, Curtis G Callan, Aleksandra M Walczak, and Thierry Mora. OLGA: fast computation of generation probabilities of B- and t-cell receptor amino acid sequences and motifs. *Bioinformatics*, 35(17):2974–2981, September 2019.

[21] M Gellert. DNA double-strand breaks and hairpins in V(D)J recombination. *Semin. Immunol.*, 6(3):125–130, June 1994.

[22] S D Fugmann, A I Lee, P E Shockett, I J Villey, and D G Schatz. The RAG proteins and V(D)J recombination: complexes, ends, and transposition. *Annu. Rev. Immunol.*, 18(1):495–527, 2000.

[23] David G Schatz and Patrick C Swanson. V(D)J recombination: mechanisms of initiation. *Annu. Rev. Genet.*, 45(1):167–202, August 2011.

[24] M Weigert, L Gatmaitan, E Loh, J Schilling, and L Hood. Rearrangement of genetic information may produce immunoglobulin diversity. *Nature*, 276(5690):785–790, 1978.

[25] B Nadel and A J Feeney. Influence of coding-end sequence on coding-end processing in V(D)J recombination. *J. Immunol.*, 155(9):4322–4329, November 1995.

[26] G H Gauss and M R Lieber. Mechanistic constraints on diversity in human V(D)J recombination. *Mol. Cell. Biol.*, 16(1):258–269, January 1996.

[27] B Nadel and A J Feeney. Nucleotide deletion and P addition in V(D)J recombination: a determinant role of the coding-end sequence. *Mol. Cell. Biol.*, 17(7):3768–3778, July 1997.

[28] D Moshous, 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. Artemis, a novel DNA double-strand break repair/V(D)J recombination protein, is mutated in human severe combined immune deficiency. *Cell*, 105(2):177–186, April 2001.

[29] Yunmei Ma, Ulrich Pannicke, Klaus Schwarz, and Michael R Lieber. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell*, 108(6):781–794, March 2002.

[30] Katherine J L Jackson, Bruno Gaeta, William Sewell, and Andrew M Collins. Exonuclease activity and P nucleotide addition in the generation of the expressed immunoglobulin repertoire. *BMC Immunol.*, 5:19, September 2004.

[31] Haihui Lu, Klaus Schwarz, and Michael R Lieber. Extent to which hairpin opening by the Artemis:DNA-PKcs complex can contribute to junctional diversity in V(D)J recombination. *Nucleic Acids Res.*, 35(20):6917–6923, October 2007.

[32] Bailin Zhao, Eli Rothenberg, Dale A Ramsden, and Michael R Lieber. The molecular basis and disease relevance of non-homologous DNA end joining. *Nat. Rev. Mol. Cell Biol.*, 21(12):765–781, December 2020.

[33] A J Feeney, K D Victor, K Vu, B Nadel, and R U Chukwuocha. Influence of the V(D)J recombination mechanism on the formation of the primary T and B cell repertoires. *Semin. Immunol.*, 6(3):155–163, June 1994.

[34] Jiafeng Gu, Sicong Li, Xiaoshan Zhang, Ling-Chi Wang, Doris Niewolik, Klaus Schwarz, Randy J Legerski, Ebrahim Zandi, and Michael R Lieber. DNA-PKcs regulates a single-stranded DNA endonuclease activity of artemis. *DNA Repair (Amst.)*, 9 (4):429–437, April 2010.

[35] S Kallenbach, N Doyen, M Fanton d'Andon, and F Rougeon. Three lymphoid-specific factors account for all junctional diversity characteristic of somatic assembly of t-cell receptor and immunoglobulin genes. *Proc. Natl. Acad. Sci. U. S. A.*, 89(7):2799–2803, April 1992.

[36] S Gilfillan, A Dierich, M Lemeur, C Benoist, and D Mathis. Mice lacking TdT: mature animals with an immature lymphocyte repertoire. *Science*, 261(5125):1175–1178, August 1993.

[37] T Komori, A Okada, V Stewart, and F W Alt. Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes. *Science*, 261(5125):1171–1175, August 1993.

[38] Eilon Sharon, Leah V Sibener, Alexis Battle, Hunter B Fraser, K Christopher Garcia, and Jonathan K Pritchard. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat. Genet.*, 48(9):995–1002, September 2016.

[39] Kai Gao, Lingyan Chen, Yuanwei Zhang, Yi Zhao, Ziyun Wan, Jinghua Wu, Liya Lin, Yashu Kuang, Jinhua Lu, Xiuqing Zhang, Lei Tian, Xiao Liu, and Xiu Qiu. Germline-encoded TCR-MHC contacts promote TCR V gene bias in umbilical cord blood T cell repertoire. *Front. Immunol.*, 10:2064, August 2019.

[40] Florian Rubelt, Christopher R Bolen, Helen M McGuire, Jason A Vander Heiden, Daniel Gadala-Maria, Mikhail Levin, Ghia M Euskirchen, Murad R Mamedov, Gary E Swan, Cornelia L Dekker, Lindsay G Cowell, Steven H Kleinstein, and Mark M Davis. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells, 2016.

[41] Mikhail V Pogorelyy, Anastasia A Minervina, Maximilian Puelma Touzel, Anastasiia L Sycheva, Ekaterina A Komech, Elena I Kovalenko, Galina G Karganova, Evgeniy S Egorov, Alexander Yu Komkov, Dmitriy M Chudakov, Ilgar Z Mamedov, Thierry Mora, Aleksandra M Walczak, and Yuri B Lebedev. Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proc. Natl. Acad. Sci. U. S. A.*, 115(50):12704–12709, December 2018.

[42] Rachel M Gerstein and Michael R Lieber. Extent to which homology can constrain coding exon junctional diversity in V(D)J recombination, 1993.

[43] N V Boubnov, Z P Wills, and D T Weaver. V(D)J recombination coding junction formation without DNA homology: processing of coding termini, 1993.

[44] Yunmei Ma, Haihui Lu, Brigette Tippin, Myron F Goodman, Noriko Shimazaki, Osamu Koiwai, Chih-Lin Hsieh, Klaus Schwarz, and Michael R Lieber. A biochemically defined system for mammalian nonhomologous DNA end joining. *Mol. Cell*, 16(5): 701–713, December 2004.

[45] Nicholas R Pannunzio, Sicong Li, Go Watanabe, and Michael R Lieber. Nonhomologous end joining often uses microhomology: implications for alternative end joining. *DNA Repair*, 17:74–80, May 2014.

[46] Howard H Y Chang, Go Watanabe, Christina A Gerodimos, Takashi Ochi, Tom L Blundell, Stephen P Jackson, and Michael R Lieber. Different DNA end configurations dictate which NHEJ components are most important for joining efficiency. *J. Biol. Chem.*, 291(47):24377–24389, November 2016.

[47] Santosh K Srivastava and Harlan S Robins. Palindromic nucleotide analysis in human T cell receptor rearrangements. *PLoS One*, 7(12):e52250, December 2012.

[48] A W Goldrath and M J Bevan. Selecting and maintaining a diverse t-cell repertoire. *Nature*, 402(6759):255–262, November 1999.

[49] Paul G Thomas and Jeremy Chase Crawford. Selected before selection: A case for inherent antigen bias in the T cell receptor repertoire. *Curr. Opin. Syst. Biol.*, 18: 36–43, December 2019.

[50] Ryan O Emerson, William S DeWitt, Marissa Vignali, Jenna Gravley, Joyce K Hu, Edward J Osborne, Cindy Desmarais, Mark Klinger, Christopher S Carlson, John A Hansen, Mark Rieder, and Harlan S Robins. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet*, 49(5):659–665, May 2017.

[51] Chirag Krishna, Diego Chowell, Mithat Gönen, Yuval Elhanati, and Timothy A Chan. Genetic and environmental determinants of human TCR repertoire diversity. *Immun. Ageing*, 17(1), December 2020.

[52] William S DeWitt, 3rd, Anajane Smith, Gary Schoch, John A Hansen, Frederick A Matsen, 4th, and Philip Bradley. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife*, 7, August 2018.

[53] Ivan V Zvyagin, Mikhail V Pogorelyy, Marina E Ivanova, Ekaterina A Komech, Mikhail Shugay, Dmitry A Bolotin, Andrey A Shelenkov, Alexey A Kurnosov, Dmitriy B Staroverov, Dmitriy M Chudakov, Yuri B Lebedev, and Ilgar Z Mamedov. Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 111(16):5980–5985, April 2014.

[54] Qian Qi, Mary M Cavanagh, Sabine Le Saux, Hong NamKoong, Chulwoo Kim, Emerson Turgano, Yi Liu, Chen Wang, Sally Mackey, Gary E Swan, Cornelia L Dekker, Richard A Olshen, Scott D Boyd, Cornelia M Weyand, Lu Tian, and Jörg J Goronzy. Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination. *Sci. Transl. Med.*, 8(332):332ra46, March 2016.

[55] Hidetaka Tanno, Timothy M Gould, Jonathan R McDaniel, Wenqiang Cao, Yuri Tanno, Russell E Durrett, Daechan Park, Steven J Cate, William H Hildebrand, Cornelia L Dekker, Lu Tian, Cornelia M Weyand, George Georgiou, and Jörg J Goronzy. Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins. *Proc. Natl. Acad. Sci. U. S. A.*, 117(1):532–540, January 2020.

[56] Sebastian Fischer, Frauke Stanke, and Burkhard Tümmler. VJ segment usage of TCR-Beta repertoire in monozygotic cystic fibrosis twins. *Front. Immunol.*, 12:599133, February 2021.

[57] Paul J Martin, David M Levine, Barry E Storer, Sarah C Nelson, Xinyuan Dong, and John A Hansen. Recipient and donor genetic variants associated with mortality after allogeneic hematopoietic cell transplantation. *Blood Adv*, 4(14):3224–3233, July 2020.

[58] Matthew P Conomos, Michael B Miller, and Timothy A Thornton. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.*, 39(4):276–293, May 2015.

[59] R Witzgall, E O'Leary, A Leaf, D Onaldi, and J V Bonventre. The krüppel-associated box-a (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. *Proc. Natl. Acad. Sci. U. S. A.*, 91(10):4514–4518, May 1994.

[60] E M Oltz. Regulation of antigen receptor gene assembly in lymphocytes. *Immunol. Res.*, 23(2-3):121–133, 2001.

[61] Kenneth Murphy and Casey Weaver. *Janeway's Immunobiology*. Garland Science, March 2016.

[62] Ivana Mikocziova, Victor Greiff, and Ludvig M Sollid. Immunoglobulin germline gene variation and its impact on human disease. *Genes Immun.*, pages 1–13, June 2021.

[63] Sicong Li, Howard H Chang, Doris Niewolik, Michael P Hedrick, Anthony B Pinkerton, Christian A Hassig, Klaus Schwarz, and Michael R Lieber. Evidence that the DNA endonuclease ARTEMIS also has intrinsic 5'-exonuclease activity. *J. Biol. Chem.*, 289 (11):7825–7834, March 2014.

[64] C T Watson and F Breden. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes & Immunity*, 13(5):363–373, May 2012.

[65] Aviv Omer, Ayelet Peres, Oscar L Rodriguez, Corey T Watson, William Lees, Pazit Polak, Andrew M Collins, and Gur Yaari. T cell receptor beta germline variability is revealed by inference from repertoire data. *Genome Med.*, 14(1):2, January 2022.

[66] Véronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.*, 33(Database issue):D256–61, January 2005.

[67] Pradyot Dash, Andrew J Fiore-Gartland, Tomer Hertz, George C Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, E Bridie Clemens, Thi H O Nguyen, Katherine Kedzierska, Nicole L La Gruta, Philip Bradley, and Paul G Thomas. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661):89–93, July 2017.

[68] Sophia Ng, Roger Lopez, Guillermina Kuan, Lionel Gresh, Angel Balmaseda, Eva Harris, and Aubree Gordon. The timeline of influenza virus shedding in children and adults in a household transmission study of influenza in managua, nicaragua. *Pediatr. Infect. Dis. J.*, 2016.

[69] Evgeny S Egorov, Ekaterina M Merzlyak, Andrew A Shelenkov, Olga V Britanova, George V Sharonov, Dmitriy B Staroverov, Dmitriy A Bolotin, Alexey N Davydov, Ekaterina Barsova, Yuriy B Lebedev, Mikhail Shugay, and Dmitriy M Chudakov. Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *J. Immunol.*, 194(12):6155–6163, June 2015.

[70] Mikhail Shugay, Olga V Britanova, Ekaterina M Merzlyak, Maria A Turchaninova, Ilgar Z Mamedov, Timur R Tuganbaev, Dmitriy A Bolotin, Dmitry B Staroverov, Ekaterina V Putintseva, Karla Plevova, Carsten Linnemann, Dmitriy Shagin, Sarka Pospisilova, Sergey Lukyanov, Ton N Schumacher, and Dmitriy M Chudakov. Towards error-free profiling of immune repertoires. *Nat. Methods*, 11(6):653–655, June 2014.

[71] Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, August 1999.

[72] Matthew P Conomos, Cecelia A Laurie, Adrienne M Stilp, Stephanie M Gogarten, Caitlin P McHugh, Sarah C Nelson, Tamar Sofer, Lindsay Fernández-Rhodes, Anne E Justice, Mariaelisa Graff, Kristin L Young, Amanda A Seyerle, Christy L Avery, Kent D Taylor, Jerome I Rotter, Gregory A Talavera, Martha L Daviglus, Sylvia Wassertheil-Smoller, Neil Schneiderman, Gerardo Heiss, Robert C Kaplan, Nora Franceschini, Alex P Reiner, John R Shaffer, R Graham Barr, Kathleen F Kerr, Sharon R Browning, Brian L Browning, Bruce S Weir, M Larissa Avilés-Santa, George J Papanicolaou, Thomas Lumley, Adam A Szpiro, Kari E North, Ken Rice, Timothy A Thornton, and Cathy C Laurie. Genetic diversity and association studies in US Hispanic/Latino populations: Applications in the hispanic community health Study/Study of latinos. *Am. J. Hum. Genet.*, 98(1):165–184, January 2016.

[73] Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2021. URL `https://CRAN.R-project.org/package=data.table`. R package version 1.14.0.

[74] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

[75] Microsoft Corporation and Steve Weston. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, 2020. URL `https://CRAN.R-project.org/package=doParallel`. R package version 1.0.16.

[76] Xiuwen Zheng, David Levine, Jess Shen, Stephanie Gogarten, Cathy Laurie, and Bruce Weir. A high-performance computing toolset for relatedness and principal component analysis of snp data. *Bioinformatics*, 28(24):3326–3328, 2012. doi: 10.1093/bioinformatics/bts606.

[77] Stephanie M. Gogarten, Tushar Bhangale, Matthew P. Conomos, Cecelia A. Laurie, Caitlin P. McHugh, Ian Painter, Xiuwen Zheng, David R. Crosslin, David Levine, Thomas Lumley, Sarah C. Nelson, Kenneth Rice, Jess Shen, Rohit Swarnkar, Bruce S. Weir, and Cathy C. Laurie. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, 28(24):3329–3331, 2012. doi: 10.1093/bioinformatics/bts610.

[78] Stephanie M. Gogarten, Tamar Sofer, Han Chen, Chaoyu Yu, Jennifer A. Brody, Timothy A. Thornton, Kenneth M. Rice, and Matthew P. Conomos. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*, 2019. doi: 10.1093/bioinformatics/btz567.

[79] Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, 2020. URL `https://CRAN.R-project.org/package=cowplot`. R package version 1.1.1.

[80] Howard H Y Chang, Go Watanabe, and Michael R Lieber. Unifying the DNA end-processing roles of the artemis nuclease. *J. Biol. Chem.*, 290(40):24036–24050, October 2015.

[81] Howard H Y Chang and Michael R Lieber. Structure-Specific nuclease activities of artemis and the artemis: DNA-PKcs complex. *Nucleic Acids Res.*, 44(11):4991–4997, June 2016.

[82] Albert G Tsai, Aaron E Engelhart, Ma'mon M Hatmal, Sabrina I Houston, Nicholas V Hud, Ian S Haworth, and Michael R Lieber. Conformational variants of duplex DNA

correlated with cytosine-rich chromosomal fragile sites. *J. Biol. Chem.*, 284(11):7157–7164, March 2009.

[83] Yunmei Ma, Klaus Schwarz, and Michael R Lieber. The Artemis:DNA-PKcs endonuclease cleaves DNA loops, flaps, and gaps, 2005.

[84] Yuliana Yosaatmadja, Hannah T Baddock, Joseph A Newman, Marcin Bielinski, Angeline E Gavard, Shubhashish M M Mukhopadhyay, Adam A Dannerfjord, Christopher J Schofield, Peter J McHugh, and Opher Gileadi. Structural and mechanistic insights into the artemis endonuclease and strategies for its inhibition. *Nucleic Acids Res.*, 49 (16):9310–9326, September 2021.

[85] Zbigniew Dominski. Nucleases of the metallo-beta-lactamase family and their role in DNA and RNA metabolism. *Crit. Rev. Biochem. Mol. Biol.*, 42(2):67–93, March 2007.

[86] Ilaria Pettinati, Jürgen Brem, Sook Y Lee, Peter J McHugh, and Christopher J Schofield. The chemical biology of human Metallo-β-Lactamase fold proteins. *Trends Biochem. Sci.*, 41(4):338–355, April 2016.

[87] Haihui Lu, Noriko Shimazaki, Prafulla Raval, Jiafeng Gu, Go Watanabe, Klaus Schwarz, Patrick C Swanson, and Michael R Lieber. A biochemically defined system for coding joint formation in V(D)J recombination. *Mol. Cell*, 31(4):485–497, August 2008.

[88] Christina A Gerodimos, Howard H Y Chang, Go Watanabe, and Michael R Lieber. Effects of DNA end configuration on XRCC4-DNA ligase IV and its stimulation of artemis activity. *J. Biol. Chem.*, 292(34):13914–13924, August 2017.

[89] I B Rogozin and N A Kolchanov. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochimica et Biophysica Acta*, 1171(1):11–18, 15 November 1992. ISSN 0006-3002. URL `http://www.ncbi.nlm.nih.gov/pubmed/1420357`.

[90] D K Dunn-Walters, A Dogan, L Boursier, C M MacDonald, and J Spencer. Base-specific sequences that bias somatic hypermutation deduced by analysis of out-of-frame human IgVH genes. *The Journal of Immunology*, 160(5):2360–2364, 1 March 1998. ISSN 0022-1767. URL `https://www.ncbi.nlm.nih.gov/pubmed/9498777`.

[91] Reuma Magori Cohen, Steven H Kleinstein, and Yoram Louzoun. Somatic hypermutation targeting is influenced by location within the immunoglobulin V region. *Molecular Immunology*, 48(12-13):1477–1483, July 2011. ISSN 0161-5890, 1872-9142. doi: 10.1016/j.molimm.2011.04.002. URL `http://dx.doi.org/10.1016/j.molimm.2011.04.002`.

[92] Gur Yaari, Jason A Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel N H Stern, Kevin C O'Connor, David A Hafler, Uri Laserson, Francois

Vigneault, and Steven H Kleinstein. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in Immunology*, 4:358, 15 November 2013. ISSN 1664-3224. doi: 10.3389/fimmu.2013.00358. URL http://dx.doi.org/10.3389/fimmu.2013.00358.

[93] Yuval Elhanati, Zachary Sethna, Quentin Marcou, Curtis G Callan, Jr, Thierry Mora, and Aleksandra M Walczak. Inferring processes underlying B-cell repertoire diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676), 5 September 2015. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2014.0243. URL http://dx.doi.org/10.1098/rstb.2014.0243.

[94] Lirong Wei, Richard Chahwan, Shanzhi Wang, Xiaohua Wang, Phuong T Pham, Myron F Goodman, Aviv Bergman, Matthew D Scharff, and Thomas MacCarthy. Overlapping hotspots in CDRs are critical sites for V region diversification. *Proc. Natl. Acad. Sci. U. S. A.*, 112(7):E728–37, February 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1500788112. URL http://dx.doi.org/10.1073/pnas.1500788112.

[95] Ang Cui, Roberto Di Niro, Jason A Vander Heiden, Adrian W Briggs, Kris Adams, Tamara Gilbert, Kevin C O'Connor, Francois Vigneault, Mark J Shlomchik, and Steven H Kleinstein. A model of somatic hypermutation targeting in mice based on High-Throughput ig sequencing data. *J. Immunol.*, 197(9):3566–3574, November 2016. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.1502263. URL http://dx.doi.org/10.4049/jimmunol.1502263.

[96] Jean Feng, David A Shaw, Vladimir N Minin, Noah Simon, and Frederick A Matsen, IV. Survival analysis of DNA mutation motifs with penalized proportional hazards. *Ann. Appl. Stat.*, 13(2):1268–1294, June 2019. ISSN 1932-6157, 1941-7330. doi: 10.1214/18-AOAS1233. URL https://projecteuclid.org/euclid.aoas/1560758446.

[97] Natanael Spisak, Aleksandra M Walczak, and Thierry Mora. Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic Acids Res.*, 48(19):10702–10712, November 2020. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkaa825. URL http://dx.doi.org/10.1093/nar/gkaa825.

[98] Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, 41(Web Server issue):W56–62, July 2013.

[99] Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, April 2016.

[100] Davis Jose, Kausiki Datta, Neil P Johnson, and Peter H von Hippel. Spectroscopic studies of position-specific DNA "breathing" fluctuations at replication forks

and primer-template junctions. *Proceedings of the National Academy of Sciences*, 106 (11):4231–4236, 2009.

[101] Andrei Slabodkin, Maria Chernigovskaya, Ivana Mikocziova, Rahmad Akbar, Lonneke Scheffer, Milena Pavlović, Habib Bashour, Igor Snapkov, Brij Bhushan Mehta, Cédric R Weber, Jose Gutierrez-Marcos, Ludvig M Sollid, Ingrid Hobæk Haff, Geir Kjetil Sandve, Philippe A Robert, and Victor Greiff. Individualized VDJ recombination predisposes the available ig sequence space. *Genome Res.*, November 2021.

[102] H Robins and O Pearson. Normal human PBMC, deep sequencing, TCRB vs TCRG comparison. `https://clients.adaptivebiotech.com/pub/TCRB-TCRG-comparison`, April 2015. Accessed: 2022-10-25.

[103] Bryan Briney, Anne Inderbitzin, Collin Joyce, and Dennis R Burton. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, 566 (7744):393–397, February 2019.

[104] Jason A Vander Heiden, Jason A Vander Heiden, Gur Yaari, Mohamed Uduman, Joel N H Stern, Kevin C O'Connor, David A Hafler, Francois Vigneault, and Steven H Kleinstein. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires, 2014.

[105] David B Jaffe, Payam Shahi, Bruce A Adams, Ashley M Chrisman, Peter M Finnegan, Nandhini Raman, Ariel E Royall, Funien Tsai, Thomas Vollbrecht, Daniel S Reyes, N Lance Hepler, and Wyatt J McDonnell. Functional antibodies exhibit light chain coherence. *Nature*, 611(7935):352–357, November 2022.

[106] Duncan K Ralph and Frederick A Matsen, 4th. Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput. Biol.*, 12(1):e1004409, January 2016.

[107] A J Feeney. Lack of N regions in fetal and neonatal mouse immunoglobulin V-D-J junctional sequences. *J. Exp. Med.*, 172(5):1377–1390, November 1990.

[108] H Gu, I Förster, and K Rajewsky. Sequence homologies, N sequence insertion and JH gene utilization in VHDJH joining: implications for the joining mechanism and the ontogenetic timing of Ly1 B cell and B-CLL progenitor generation. *EMBO J.*, 9(7): 2133–2140, July 1990.

[109] Michael R Lieber. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.*, 79:181–211, 2010.

[110] Howard H Y Chang, Nicholas R Pannunzio, Noritaka Adachi, and Michael R Lieber. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell Biol.*, 18(8):495–506, August 2017.

[111] Nicholas R Pannunzio, Go Watanabe, and Michael R Lieber. Nonhomologous DNA end-joining for repair of DNA double-strand breaks. *J. Biol. Chem.*, 293(27):10512–10523, July 2018.

[112] Michael R Lieber. The mechanism of human nonhomologous DNA end joining. *J. Biol. Chem.*, 283(1):1–5, January 2008.

[113] Tina Funck, Mike Bogetofte Barnkob, Nanna Holm, Line Ohm-Laursen, Camilla Slot Mehlum, Sören Möller, and Torben Barington. Nucleotide composition of human Ig nontemplated regions depends on trimming of the flanking gene segments, and terminal deoxynucleotidyl transferase favors adding cytosine, not guanosine, in most VDJ rearrangements. *The Journal of Immunology*, 201(6):1765–1774, 2018.

[114] Jiafeng Gu, Haihui Lu, Brigette Tippin, Noriko Shimazaki, Myron F Goodman, and Michael R Lieber. XRCC4:DNA ligase IV can ligate incompatible DNA ends and can ligate across gaps, 2007.

[115] Jiafeng Gu, Haihui Lu, Albert G Tsai, Klaus Schwarz, and Michael R Lieber. Single-stranded DNA ligation and XLF-stimulated incompatible DNA end ligation by the XRCC4-DNA ligase IV complex: influence of terminal DNA sequence. *Nucleic Acids Res.*, 35(17):5755–5762, August 2007.

[116] Peter Ahnesorg, Philippa Smith, and Stephen P Jackson. XLF interacts with the XRCC4-DNA ligase IV complex to promote DNA nonhomologous end-joining. *Cell*, 124(2):301–313, January 2006.

[117] Takashi Ochi, Andrew N Blackford, Julia Coates, Satpal Jhujh, Shahid Mehmood, Naoka Tamura, Jon Travers, Qian Wu, Viji M Draviam, Carol V Robinson, Tom L Blundell, and Stephen P Jackson. DNA repair. PAXX, a paralog of XRCC4 and XLF, interacts with ku to promote DNA double-strand break repair. *Science*, 347(6218): 185–188, January 2015.

[118] Nelli Heikkilä, Reetta Vanhanen, Dawit A Yohannes, Iivari Kleino, Ilkka P Mattila, Jari Saramäki, and T Petteri Arstila. Human thymic T cell repertoire is imprinted with strong convergence to shared sequences. *Mol. Immunol.*, 127:112–123, November 2020.

[119] Nelli Heikkilä, Silja Sormunen, Joonatan Mattila, Taina Härkönen, Mikael Knip, Emmi-Leena Ihantola, Tuure Kinnunen, Ilkka P Mattila, Jari Saramäki, and T Petteri Arstila. Generation of self-reactive, shared T-cell receptor alpha chains in the human thymus. *J. Autoimmun.*, 119(102616):102616, May 2021.

[120] David Jung and Frederick W Alt. Unraveling V(D)J recombination. *Cell*, 116(2): 299–311, January 2004.

[121] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

[122] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *arXiv [cs.LG]*, May 2021.

[123] Yuval Elhanati, Zachary Sethna, Curtis G Callan, Jr, Thierry Mora, and Aleksandra M Walczak. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol. Rev.*, 284(1):167–179, July 2018.

[124] Mikhail V Pogorelyy, Anastasia A Minervina, Mikhail Shugay, Dmitriy M Chudakov, Yuri B Lebedev, Thierry Mora, and Aleksandra M Walczak. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol.*, 17(6): e3000314, June 2019.

[125] Jared Dean, Ryan O Emerson, Marissa Vignali, Anna M Sherwood, Mark J Rieder, Christopher S Carlson, and Harlan S Robins. Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. *Genome Med.*, 7:123, November 2015.

[126] B Devlin and K Roeder. Genomic control for association studies. *Biometrics*, 55(4): 997–1004, December 1999.

[127] Matthew L Freedman, David Reich, Kathryn L Penney, Gavin J McDonald, Andre A Mignault, Nick Patterson, Stacey B Gabriel, Eric J Topol, Jordan W Smoller, Carlos N Pato, Michele T Pato, Tracey L Petryshen, Laurence N Kolonel, Eric S Lander, Pamela Sklar, Brian Henderson, Joel N Hirschhorn, and David Altshuler. Assessing the impact of population stratification on genetic association studies. *Nat. Genet.*, 36(4):388–393, April 2004.

[128] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, 11(7): 459–463, July 2010.