On the diversity of T cell receptors in the genus *Mus*

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Moritz August Peters

aus Münster

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 16.09.2024

Dekan: Prof. Dr. Thilo Stehle

1. Berichterstatter: Prof. Dr. Yingguang Frank Chan

2. Berichterstatter: Prof. Dr. Marja Timmermans

Contents

Summary ... *1*

Zusammenfassung ... *3*

Introduction ... *5*

Structure and Function of TCRs ... 6

Generation of a diverse TCR repertoire ... 8

Selection of developing T cells in the thymus ... 13

Size estimates of TCR repertoires ... 16

Methods for TCR sequencing library preparation ... 18

Methods for TCR repertoire analysis and comparison ... 22

Mouse models in studies of the adaptive immune system diversity ... 25

Multi-omic analysis of gene regulation ... 28

Objectives ... *31*

Chapter 1: Distinct evolution at TCRα and TCRβ loci in the genus Mus ... *33*

Chapter 2: Genetic determinants of distinct CD8$^+$ α/β-TCR repertoires in the genus Mus ... *57*

Chapter 3: Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells ... *124*

Discussion ... *152*

On the co-evolution of TCRs and MHCs ... 153

The effect of MHC heterozygosity on the TCR repertoire ... 155

Sharing of TCRs – How to become public ... 157

CITR-seq can be used in various research areas – an outlook ... 162

Closing Remarks ... 164

Acknowledgments ... *165*

Glossary ... *167*

References ... *169*

Summary

The diversity of T cell receptors (TCRs) is one of the backbones of an effective adaptive immune system. This diversity is generated by somatic rearrangements of gene segments in two separate peptide chains that dimerize to form a unique receptor that can specifically recognize antigens presented by major histocompatibility complexes (MHCs). The generative process of TCR repertoire formation is largely defined by stochastic events that can theoretically give rise to more than 10^{15} unique receptors. Strikingly, immune responses to common pathogens are frequently driven by identical or very similar TCRs. Consequently, there is significant non-random sharing of such “public” receptors between individuals. This has invoked the idea that genetically encoded factors contribute to the shaping of an individual’s TCR repertoire, but experimental validation of such factors has been lacking due to the technical challenge of capturing the sheer size of diverse TCR repertoires.

Together with my colleagues, I have developed a single-cell and high-throughput TCR sequencing protocol capable of generating paired TCR sequencing data from millions of individual $CD8^+$ T cells. To reveal the contribution of genetic factors in the generation of those TCRs, we generated TCR repertoires from 32 mice representing the reference lab mouse and three sister species, as well as F1 hybrids between them. Collectively, these mice span an evolutionary divergence time of approximately three million years and represent an exceptional model to study germline determinants of TCR repertoire formation, owing to their distinct genetic backgrounds. By conducting a comprehensive comparison of the *variable*, *diversity* and *joining* gene segments across the different species, we showed that despite notable evolutionary conservation at much of the loci, the TCR alpha variable gene segment locus has undergone a major locus expansion as indicated by the significantly different number of gene segments across all species. Following this observation, we were able to show that the usage frequencies of gene segments of TCRs varied significantly across species but were remarkably conserved in intra-species repertoires. Using F1 hybrids, we can demonstrate genetic control in usage for specific gene segments, because individual parental alleles retain differential usage frequencies despite a shared heterozygous genetic background. Further we have

evaluated the impact of thymic selection on the shaping of an individual's repertoire. TCR repertoire diversity reduction caused by thymic selection is mostly defined by rejection of variable gene segments in TCR beta chains and occurs strictly through direct protein-protein interaction with antigen-presenting major histocompatibility complex alleles. This has significant consequences for the sharing of identical and similar TCRs across several individuals. We showed that public paired TCR motifs are approximately four times more frequent than expected by chance but are still extremely rare compared to the sharing of identical single-chain motifs. Further, by comparing the frequencies of short amino acid motifs from the antigen-specific region of TCRs, we show that even in those regions, arising from seemingly random fusion of gene segments, abundances of particular amino acids motifs are remarkably dependent of the respective genotype of an individual. This work not only provides an approach to analyze TCR repertoires at unprecedented scale but also reveals a surprising extent of genetic contribution to the shaping of an individual's TCR repertoire.

Zusammenfassung

Die Diversität der Rezeptoren von T Zellen (TCRs) stellt einen der wichtigsten Faktoren für das intakte adaptive Immunsystem dar. Diese Diversität wird in erster Linie durch die separate somatische Rekombination verschiedener Gensegmente in zwei Peptidketten generiert, die durch Dimerisierung einen einzigartigen Rezeptor bilden, der wiederum spezifisch die von Haupthistokompatibilitätskomplexen (MHCs) präsentierten Antigene erkennt. Der Prozess, durch den ein TCR-Repertoire generiert wird, ist größtenteils stochastischer Natur und kann potenziell bis zu 10^{15} verschiedene Rezeptoren hervorbringen. Erstaunlicherweise werden bei der individuellen Immunantwort gegen geläufige Pathogene häufig identische oder sehr ähnliche TCRs verwendet. Daraus ergibt sich die Annahme, dass solche „gebräuchlichen“ TCRs nicht auf rein zufälliger Basis generiert werden. Es ist deshalb die Hypothese entstanden, dass genetische Faktoren eine entscheidende Rolle in der Zusammensetzung eines TCR-Repertoires spielen. Aufgrund der schieren Größe von vollständigen TCR-Repertorien ist es bisher jedoch schwierig gewesen, diese Hypothese mit Hilfe von großen TCR-Datensätzen zu überprüfen. Zusammen mit meinen Kollegen habe ich ein Hochdurchsatz-Protokoll für die Analyse von mehreren Millionen einzelnen $CD8^+$ T Zellen und deren gepaarten TCRs entwickelt. Um den Einfluss genetischer Faktoren in der Entstehung dieser TCRs zu analysieren, haben wir die Methode an 32 Mäusen angewandt, die vier verschiedenen Inzuchtlinien angehören, die ursprünglich aus wilden Populationen entnommen wurden. Unter diesen Mäusen befanden sich auch F1 Hybride aus Kreuzungen mit der häufig verwendeten C57BL/6 Labormauslinie. Diese Mäuse repräsentieren gemeinsam eine evolutionäre Divergenz von etwa drei Millionen Jahren und stellen somit, dank der einheitlichen genetischen Eigenschaften, ein hervorragendes Modellsystem dar, um die vererblichen Faktoren für die Generierung eines TCR-Repertoires zu analysieren. Zunächst haben wir die Gen-Loci der sogenannten *variable*, *diversity* und *joining* Gensegmente der verschiedenen Mausarten systematisch verglichen und konnten zeigen, dass während die Mehrheit dieser Loci konserviert sind, der Gen-Lokus der variablen Gensegmente des Alpha-TCRs von umfänglichen Genduplikationen betroffen ist. Dies hat vor allem die Konsequenz, dass sich die verschiedenen Mausarten durch eine sehr unterschiedliche Anzahl an variablen Gensegmenten des Alpha-TCRs

auszeichnen. In der Folge konnten wir zeigen, dass sich die Nutzungsfrequenz der verschiedenen Gensegmente zwischen den Mausarten stark unterscheidet, jedoch innerhalb einer Art wenig variiert. Mit Hilfe der F1 Hybride konnten wir feststellen, dass die Nutzung der Gensegmente einer genetischen Kontrolle unterliegt, da wir die parentalen Nutzungsmuster auch in dem entsprechenden heterozygoten genetischen Hintergrund der F1 hybride nachweisen konnten. Darüber hinaus haben wir den Einfluss der Selektion im Thymus auf das TCR-Repertoire analysiert. Wir haben gezeigt, dass sich eine selektionsbedingte Reduktion der TCR-Diversität vor allem durch den Ausschluss einzelner variabler Gensegmente des Beta-TCRs auszeichnet und dieser Ausschluss stark vom MHC-Typ eines Individuums abhängt. Diese Beobachtung hat auch wichtige Auswirkungen in Bezug auf die Wahrscheinlichkeit einen identischen TCR in zwei Individuen vorzufinden. Wir konnten nachweisen, dass diese Wahrscheinlichkeit etwa viermal höher ist als durch Zufall erwartet, was allerdings noch immer sehr viel seltener ist als das wiederholte Auffinden einer einzelnen Alpha- oder Beta-Kette in zwei Individuen. Darüber hinaus haben wir die Häufigkeit von Aminosäuremotiven aus der antigenspezifischen Region von TCRs in den verschiedenen Mausarten verglichen. Obwohl diese Motive hauptsächlich durch stochastische Prozesse entstehen konnten wir nachweisen, dass ihre Häufigkeit in bemerkenswerter Weise vom Genotyp eines Individuums abhängen. Diese Arbeit präsentiert nicht nur ein Verfahren, mit dem sich das TCR-Repertoire in nie dagewesener Tiefe analysieren lässt, sondern zeigt auch, wie sehr sich genetische Faktoren auf die Zusammensetzung eines TCR-Repertoires auswirken.

Introduction

Selective pressure caused by the evolutionary arms race between host and infectious organisms has led to the development of various defense mechanisms across all multicellular organisms. Taking a broad view, these protective mechanisms, commonly referred to as an individual's immune system, can be divided into innate and adaptive responses. Both types of responses are necessary to distinguish self from non-self to repel pathogenic challenges while preserving self-tolerance. Typically, this is accomplished through receptor-ligand interactions, whereby extracellular stimuli are transmitted into the cell to trigger an immune response. In the case of the more evolutionarily ancient innate immune system, receptor specificity is germline-encoded and has often evolved to target invariant molecular structures of pathogens, for instance lipopolysaccharide (LPS), a major component of the outer membrane of all gram-negative bacteria. One of the main types of these pattern recognition receptors (PRRs) is known as toll-like receptors and was first discovered in Tübingen in 1985 [1]. Critically, the innate immune system is limited to a set of common, recognizable pathogenic molecular patterns, whose diversity may seem far too low in the face of vast number of pathogens present in an individual's environment. In addition, pathogens have evolved a diverse repertoire of counterstrategies to impair PRR-mediated signaling in the innate immune system (reviewed in [2]). Collectively, this has favored the evolution of a secondary defense strategy – the adaptive immune system.

An adaptive immune system can be found in all vertebrates including agnathans and it is therefore believed to have evolved roughly 500 million years ago [3, 4]. One of its key features is the presence of a dichotomic cell lineage known as lymphocytes, which consist of B and T cells that were first described in 1965 [5]. B and T cells both express diverse repertoires of adaptive immune receptors that collectively can recognize a remarkably large number of antigens. Despite the distinct roles of B and T cells in adaptive immunity, the generative process of their adaptive immune receptors is very similar. Here, I will elaborate specifically on the generation, selection and function of T cell receptors (TCRs) and provide an overview of past and present TCR repertoire analysis approaches.

Structure and Function of TCRs

T cells constantly patrol the body and scan their surroundings for pathogenic infections or aberrant cells. Recognition of these threats is facilitated by a surface bound heterodimeric receptor – the T cell receptor. Its discovery dates back to the early 1980's and TCRs have been subject of extensive research ever since [6, 7]. I will focus my discussion on the primary class of TCRs consisting of a TCRα and TCRβ chain expressed by approximately 95% of T cells (the rest being a second class of TCRs consisting of γ/δ-heterodimers). Depending on the mutually exclusive expression of the co-receptor CD4 or CD8 in the different sub-classes of T cells [8, 9], TCR recognize short peptides presented by major histocompatibility complexes (MHC) class I or II. The requirement for those short peptides (hereafter called antigens) to be presented by MHC molecules is referred to as MHC-restriction and depicts one of the main functional differences between TCRs and B cell surface receptors (BCRs) as well as their soluble form - the antibodies [10]. TCRs on the surface of $CD8^+$ T cells recognize antigens presented by MHC class I molecules that are present on all nucleated cells. Typically, these antigens consist of 8-10 amino acid residues [11, 12] and are generated by proteasomal degradation of intracellular proteins. Critically, it has been shown that MHC class I molecules can also be loaded with peptides derived from extracellular proteins in a process called *cross-priming*, which is pivotal for the defense against tumors and viruses [13]. Activation of $CD8^+$ T cells by recognition of a foreign antigen results in the release of two cytotoxic molecules: granzyme B and perforin, which in turn trigger apoptosis in the recognized infected or aberrant target cell [14]. In contrast to this, TCRs on CD4+ T cells recognize antigens presented by MHC class II complexes that are expressed on antigen presenting cells (APCs, such as B cells and dendritic cells). These antigens are slightly longer peptides (approximately 13-25 amino acids [15]) and critically, emerge from degradation of endocytosed extracellular proteins. Activated CD4+ T cells secrete various cytokines which in turn can activate cells of the innate immune system and fine-tune ongoing immune responses (reviewed here [16]). The collective set of antigens presented by both MHCs is referred to as the immunopeptidome and largely depends on the MHC haplotype of an individual. In humans, the human-leukocyte antigen (HLA, human MHC) locus is considered to be the most diverse region in the entire genome with several tens of

thousands of identified haplotypes [17-19]. Genetic MHC variation has been shown to directly shape the TCR repertoire [20, 21]. This phenomenon is believed to be primarily driven by the distinct affinity characteristics exhibited by a given TCR and the MHC molecules specific to the underlying MHC haplotype, a topic that will be further addressed in subsequent sections. The focal contact regions of a TCR to the MHC complex are three complementarity determining regions (CDR1-CDR3). CDR1 and CDR2 are germline-encoded short sequences and polymorphisms in these regions have been shown to modify the TCRs affinity to MHC complexes [22, 23]. This led to the conclusion that CDR1 and CDR2 are most critical for MHC-TCR contact maintenance rather than antigen recognition. CDR3 is the most diverse sequence and unique to every T cell because it consists of the junctional regions resulting from the somatic recombination of gene segments (described below). Crystal structures of TCR-MHC complexes have shown that the residues within the CDR3 region are in closest proximity to the MHC bound antigen [24-26]. Because of its immense diversity across TCRs and the close proximity to the antigen in TCR-MHC complexes, CDR3 sequences are therefore believed to primarily define the antigen specificity of the underlying TCR.

The origin of MHC restriction in TCR antigen recognition remains an intensely debated topic with two opposing principal models. The *germline model* states that the ability of TCRs to bind to MHCs is germline-encoded and the amino acid residues that mediate this interaction (e.g., in CDR1 and CDR2) are conserved and have co-evolved between TCRs and MHCs [27]. This hypothesis is supported by multiple lines of evidence. For instance, the topology of many published TCR-MHC structures exhibits remarkable conservation [28, 29] and there is evidence for particular germline-encoded amino acid residues of TCRs that mediate MHC contact [30]. On the other hand, it has been demonstrated that both the identity of the antigen [31] as well as the identity of the CDR3 sequence [32] can significantly alter the contact sites at which the TCR engages the MHC. Further, examples of autoreactive TCR that recognize self-peptide MHC complexes were shown to utilize uncommon MHC contact sites [33]. An important finding that conflicts with the germline hypothesis was that T cells that lack the germline-encoded CDR1 and CDR2 sequences maintain full functionality, including the engagement of MHCs [34].

The second alternative model to explain MHC restriction of TCRs is the *co-receptor hypothesis*. It has been shown that the lineage marker co-receptors CD4 and CD8 also bind the MHC-TCR complex with the main task of recruiting the receptor tyrosine kinase *Lck* [35]. TCRs also form complexes with CD3 molecules, which play a crucial role in transmitting TCR signaling in the cytosolic portion of the complex through phosphorylation cascades [36]. The full assembly of these complexes leads to a phosphorylation cascade that is required to initiate an immune response. The *co-receptor hypothesis* therefore states that the orchestration of signaling at the immunological synapse, which necessitates MHC engagement by CD4 or CD8, imposes MHC restriction on the TCR. In support of this hypothesis, it was shown that mice lacking both CD4/CD8 and MHC-I/MHC-II can still generate functional TCRs that can recognize specific epitopes [37]. However, the diversity of epitopes of such MHC-independent TCRs has yet to be shown to resemble the diversity observed in general pre-selection TCR repertoires. In summary, neither hypothesis has been convincingly rejected nor definitively proven to be correct thus far.

Generation of a diverse TCR repertoire

In 1957, Frank Macfarlane Burnet published a paper that introduced the *clonal selection theory* as possible explanation for the flexibility and diversity within the adaptive immune system [38]. In the following 20 years, evidence started to accumulate that the key to generation of diversity in TCRs (and B cell receptors) is their generation by somatic rearrangements of multiple gene segments [39]. The underlying mechanisms, known as V(D)J recombination was first described in 1976 [40] and explained the long-standing question of how millions of unique antigen receptors could be generated from a set of roughly 20,000 genes in human. The term *V(D)J recombination* relates to the underlying gene segments that are recombined to generate the heterodimeric α- and β-chains (or γ- and δ-chains) of the mature TCR. These segments consist of variable (V), diversity (D, exclusive to β-chains) and joining (J) gene segment distributed across hundreds of kb in the genomes of mice (chromosome 6 and 14) and human (chromosome 7 and 14) [41] (see **Fig. 1**). Comparative genomics of TCR loci across various vertebrate species has unveiled significant differences in the absolute numbers of individual gene segment [42-

44] (elaborated later). Furthermore, many of the identified gene segments exhibit substantial sequence identity ranging from 70-100%, which provides evidence for their generation by means of gene duplication events.

Somatic rearrangement of individual gene segments requires the precise execution of an ordered series of DNA double-strand breaks and subsequent DNA repair mechanisms. This intricate process is mediated by an enzymatic complex of two DNA recombinases, known as recombination activating genes-1 and -2 (*Rag-1* and *Rag-2*) [45]. The absence of one of these two genes has been demonstrated to completely impair V(D)J recombination, leading to the arrests of T and B cell development in mice [46, 47]. Considering that millions of T cells perform V(D)J recombination on a daily basis and that it involves introducing double-strand breaks to DNA, expression of *Rag1* and *Rag2* needs to be extremely tightly regulated and highly cell- and developmental timing-specific. Otherwise, it can lead to highly deleterious outcomes. Indeed, it has been demonstrated that ubiquitous *Rag1/2* expression causes severe phenotypes in mice [48]. Sequence analysis of Rag proteins indicates that they originate from transposons but have almost completely lost their transposase activity in favor of acquiring the function of a recombinase [49]. The *Rag1/Rag2* complex (referred to as Rag-complex from now on) specifically targets recombination signal sequences (RSS) that are located between every V(D)J gene segment [50]. RSSs consist of a conserved heptamer sequence, a spacer sequence with a conserved length of either 12 or 23 base pairs and a conserved nonamer sequence. Initially, the Rag complex binds either a 12- or 23-RSS and subsequently has a strong preference to bind and cleave a second RSS with the respective alternative spacer length [51]. This specific preference of spacer length combinations is known as the 12/23 rule and ensures that recombination only occurs between segments of different spacer lengths. Accordingly, V and J gene-segments are flanked by RSSs with spacers of identical length with additional mechanisms in place to ensure integration of a D segment in TCRβ-chains [52]. Consequently, implementation of the 12/23 rule ensures that V(D)J recombination results in the fusion of a single V to (D) to J segment. The sequence identity of the RSS has been shown to impact the recombination efficiency, possibly through modulating the Rag-complex binding strength and thereby alter the usage frequencies of particular V(D)J segments [53].

Double-strand breaks introduced by the Rag-complex have been shown to be repaired by non-homologous end joining (NHEJ) [54]. Importantly, NHEJ is highly mutagenic, in that the ligation of accessible DNA coding ends is imprecise, leading to non-template insertions mediated by terminal deoxynucleotidyl transferase (TdT) [55]. Nucleotides at these junctional DNA overhangs can also be removed prior to final segment ligation, a process that is not yet fully understood. Enzymes possessing endonuclease activity (e.g., Artemis) and are involved in double-strand break repair have been shown to be involved in nucleotide deletion at the Rag recombination sites [56]. In addition to the combinatorial recombination of gene segments, stochastic nucleotide insertions and deletions serve as the primary source of TCR diversity. However, because of a lack of control over the number of insertions and deletions at each junction site, they pose a high risk of introducing frame shifts in the resulting TCR transcript. In theory, only 1/3 of TCRs should remain in-frame following the deletion or addition of nucleotides. In practice, out-of-frame TCRs are frequently observed in TCR datasets however, the reported frequencies vary considerably [57]. A potential reason for this variation is the effective degradation of these non-functional TCRs by nonsense-mediated decay, making their detection dependent on the type of method and its sensitivity [58]. Regardless of their precise frequency, out-of-frame TCRs are often seen as “passengers” within T cells that ultimately rearranged a functional TCR from their other allele. As such, they do not undergo any TCR specificity-driven selection and can therefore be used to compare V(D)J diversity pre and post thymic selection [59, 60].

V(D)J recombination at the TCRα and TCRβ locus occurs in a sequential, stepwise fashion, with TCRβ recombination occurring first. Within each individual locus, chromatin modifications have been shown to specifically modulate the accessibility of RSS sequences and thereby control the order of recombination of V, D and J segments [61, 62]. It is currently believed that *cis*-acting promoters are guiding the recruitment of chromatin remodelers to remove nucleosomes from RSS sequences, which in turn makes them accessible for the Rag-complex [63]. One of such enhancers is the Eβ enhancer that has been shown to modulate chromatin accessibility specifically in the cluster of D- and J-segments [61, 64]. Consequently, in the TCRβ locus D-to-J joining precedes V-to-DJ joining [65]. Despite the presence of two TCRβ alleles in the genome, each individual

T cell expresses just one cell-specific TCR. Accordingly, rearrangements and subsequent expression needs to be repressed on one of the two alleles, a process known as allelic exclusion. Once an in-frame TCRβ chains has been successfully rearranged from one allele, feedback inhibition inhibits further rearrangements on the respective other allele. The nuclear localization of alleles has been demonstrated to be of critical importance for mono-allelic initiation of V(D)J recombination [66]. Despite this inhibition by spatial localization it was also shown that TCRβ-rearrangements occur in an asynchronous fashion on both alleles [65]. Collectively, several mechanisms are in place to ensure precisely timed rearrangements and allelic exclusion in the TCRβ locus some of which remain to be further characterized.

In contrast to that, rearrangements within the TCRα locus happen simultaneously on both alleles in a continuous fashion without strictly enforced allelic exclusion [67]. Continuous rearrangements of the locus lead to biases of V-J gene usage based on their respective location with initial recombination events preferentially incorporating 3' Vα segments and 5' Jα segments [68]. Different promotors are involved in the initial and late recombination events that have been shown to have a distinct target range of Jα and Vα segments [69, 70]. Collectively, this less tightly regulated V(D)J recombination enables multiple testing of TCRα chains in combination with the previously fixed mono-allelic expressed TCRβ chain during thymic selection of the assembled paired TCR [71, 72]. A relevant side-effect of the continuous rearrangement of both TCRα is the increased frequency of T cells that express two in-frame α-chains [73] with severe implications for autoimmune reactions [74, 75]. Mice with transgenic fluorescent-TCRα reporters were used to show that approximately 16% of T cells express two functional TCRα chains [76].

Apart from a potent recombination machinery, somatically rearranged immune receptors depend on the presence of numerous gene segments, that are combinatorically fused to encode a unique receptor. The number of available gene segments for V(D)J recombination has been shown to vary immensely across different species [43, 87]. The individual gene segments have been multiplied by varying extents of gene duplication. For instance, comparative genomics of Vα and Vβ gene segments has provided evidence that all current V gene segments originate from five ancestral Vα and four ancestral Vβ genes [87]. The expansions (and contractions) of the V(D)J gene segment loci are

consistent with the postulated hypothesis of birth-and-death of multigene families [88]. It is derived from the observation that intra-individual gene segment sequence identity is not necessarily larger than inter-species gene segment sequence identity. For example, the homology of TCR V gene families between mice and humans has been shown to be larger than the sequence identity across families within both species [89]. In the murine TCR Vα cluster, a relatively recent major duplication of roughly two-thirds of all gene segments has been described [41]. Little is known about the effects of those major rearrangements on TCR repertoire diversity. In chapters 1 and 2 of this thesis, we performed a detailed analysis of the TCR locus structure of wild-derived inbred mouse species (introduced later) as well as the associated diversity variance in their TCR repertoires.

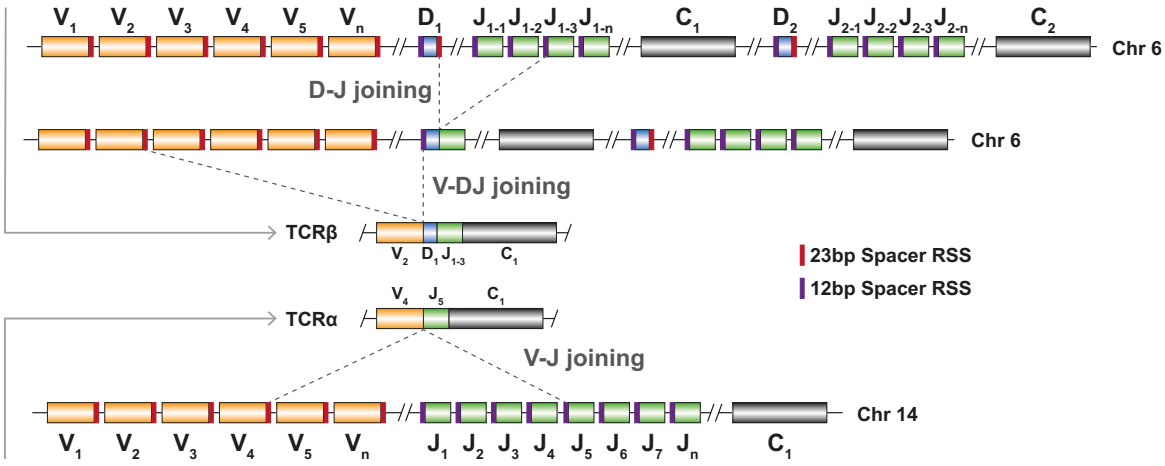


Figure 1: Schematic of V(D)J recombination. *The TCRβ locus (**top**) contains one cluster of V gene segments and two clusters of J gene segments, each with a respective D gene segment and a constant region. TCRβ rearrangement is initiated by D-J joining, guided by the Rag-complex through recognizing a 23-RSS motif at the 3' end of a D segment and a 12-RSS motif at the 5' end of a J segment. Subsequently, the DJ-sequence is joined to a V gene segment containing a 23-RSS motif at the 3' end. The TCRα locus (**bottom**) only contains a cluster of V gene segments and a cluster of J gene segments alongside a constant region. In contrast to the TCRβ locus, the V and J segments of the TCRα locus can continuously rearrange until an in-frame TCR chain is generated.*

Selection of developing T cells in the thymus

TCRs must meet two fundamental requirements for the maintenance of their specific functions elaborated above. Firstly, their antigen recognition ability needs to be strictly limited to “foreign” antigens to prevent immune reactions directed against the host’s healthy tissue. Secondly, TCRs need to be able to recognize and bind to MHC complexes [90] due to their MHC-restricted nature of antigen recognition. Failure in one of these abilities can result in autoimmunity in the former case or immunodeficiency in the latter. The TCR of each developing T cell is tested for these requirements during the maturation period in the thymus. A two-step process of *positive* and *negative selection* ensures that TCRs of mature T cells exhibit very defined antigen binding characteristics (see **Fig. 2**). In the murine thymus a remarkably high number of up to 50 million developing T cells undergo these selection steps on a daily basis and about 95% do not survive the process [91]. In humans the rates of cells undergoing thymic selection varies significantly across an individual’s lifespan exhibiting a gradual decline with age [92, 93]. Decreased rates of T cell selection are accompanied by degeneration of the thymus known as thymic involution, which is believed to be one of the main causes of increased disease susceptibility with age [94]. In aged human individuals, T cell homeostasis is primarily maintained by proliferation of peripheral T cells rather than thymic output [95]. In contrast to that murine thymuses sustain a life-long production of new naïve T cells [96]. The thymic output of T cells can directly be measured by quantifying T cell receptor excision circles (TRECs). These short circular DNA sequences are byproducts of V(D)J recombination arising from the excision of DNA sequences in-between gene segments. Because of their high stability but incapability to multiply they are diluted in proliferating peripheral T cells. Consequently, high levels of TRECs are used to classify T cells as recent thymic emigrants [97, 98].

After migrating to the thymus, T cells undergo a characteristic developmental program by migrating through different areas within the organ. A marker associated with the earliest stages of intra-thymic T cell development and restriction of multipotent progenitors to the T cell lineage is the expression of *Notch1* [99]. All subsequent maturation stages are typically classified by the expression of the linage markers CD4 and CD8. Initially T cells do not express any of the two markers (double negative, DN1-DN4 stages), then become

double positive (DP) before eventually showing mutually exclusive expression of one of the two markers (CD8 single positive or CD4 single positive) [100, 101]. TCRβ rearrangements are initiated at the DN3 stage of T cell development and the successful rearrangement of a TCRβ chain is required for progression beyond the β-selection checkpoint. This was initially shown by the reversal of developmental arrests in *Rag*-deficient mice upon expression of a transgenic rearranged TCRβ chain [102]. Passing of the β-selection checkpoint inhibits secondary rearrangements of the TCRβ locus, initiates expression of CD4 and CD8 and promotes rearrangements in the TCRα locus [103]. The rearranged TCRβ initially assembles in a pre-TCR complex with a pre-TCRα-chain and CD3 molecules [104]. This pre-complex is thought to prevent premature degradation of the TCRβ chain prior to the complete assembly of the full TCR. Eventually, the pre-TCRα-chain is replaced by a fully rearranged TCRα-chain and the fully assembled TCR can subsequently be subject to positive selection. Positive selection is orchestrated by cortical epithelial cells (cTECs) that load their MHC-I complexes using peptides generated by a proteasome that has a unique β5t subunit [105]. Similarly, MHC-II complexes are loaded with peptides that are also produced by a thymus specific protease (cathepsin L and TSSP) [106, 107]. Consequently, the peptidome utilized for positive selection by cTECs consists of a unique set of peptides that differs from those presented on extra-thymal MHCs. Mounting evidence now suggests that the mTEC specific presented peptidome consists of less hydrophobic peptides which in turn might lead to reduced TCR-MHC binding strength during positive selection [108]. Additionally, utilizing a unique peptidome in this initial selection step ensures that a selected TCR does not encounter identical peptides in subsequent selection steps. In a period of three to four days assembled TCRs can audition several times to be positively selected for their ability to sufficiently bind MHCs. Within this timeframe, TCRβ chains can be paired with multiple different TCRα-chains resulting from the continuous rearrangement of the locus. In this initial testing phase premature apoptosis is prevented by the gene *Bclx* [109]. Afterwards incapability to recognize MHCs results in a process called "*death by neglect*". The small fraction of T cells that show appropriate self-MHC affinity, progress to migrate to the medulla. This relocation is mediated by chemotaxis with T cells initiating *Ccr7* expression while medullary thymic epithelial cells (mTECs) express the corresponding ligands *Ccl19* and

Ccl21 [110]. It has been estimated that approximately $5x10^5$ T cells undergo negative selection in the murine thymus each day [111]. mTECs express a remarkably broad range of otherwise tissue specific antigens that seem to be unnecessary for mTEC functionality [112]. The expression of such tissue specific antigens is mediated by the transcriptional regulator *Aire* which was originally discovered as a gene involved in a severe autoimmune phenotype [113]. *Aire* binds to repressive elements, removes the repressive marks and thereby allows the expression of the underlying genes. Interestingly mTECs “hand-over” their antigens to APCs such as dendritic cells, which are then crucially involved in the negative selection process [114-116]. TCRs that violate tolerance to self-antigens undergo apoptosis initiated by the pro-apoptotic gene *Bim* that can overwrite the survival signals provided by *Bclx* [117]. The few T cells that ultimately survive both, positive and negative selection then finally undergo metabolic changes leading to the ability of rapid clonal expansion instead of induction of apoptosis after strong antigen engagement in the periphery [118]. Collectively, the outlined mechanisms ensure that the TCR repertoire consists of a diverse set of TCRs that will exclusively initiate immune responses following the detection of foreign antigens.

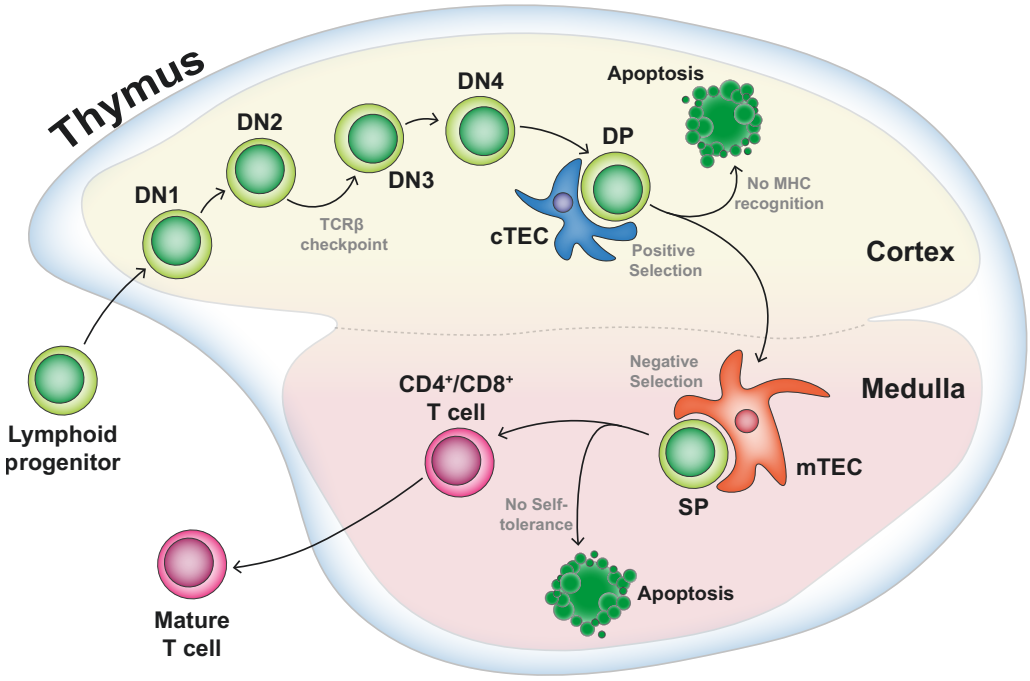


Figure 2: Schematic representation of T cell maturation and thymic selection. *Lymphoid progenitors migrate to the thymic cortex and initially progress through four double negative (DN) stages. At the transition from DN2 to DN3 the successful rearrangement of a functional TCRβ chain is evaluated. Subsequently, expression of CD4 and CD8 is initiated (double positive, DP stage) and the assembled TCR is tested for its ability to bind to self-MHCs on the surface of cortical thymic epithelial cells (cTCEs) during positive selection. T cells that show no adequate MHC affinity undergo apoptosis. Afterwards, T cells commit to either the CD4 or CD8 lineage and become single positive (SP) for these markers. Selected T cells migrate to the Medulla, where they are engaged by medullary thymic epithelial cells (mTECs) that express a different set of self-MHCs on their surface. T cells that show strong affinity to self-MHCs undergo apoptosis. Alternatively, self-tolerant T cells finish the maturation process and are released from the thymus.*

Size estimates of TCR repertoires

TCR diversity is mainly established through three different mechanisms during T cell maturation: 1) the somatic rearrangement of a diverse set of gene segments during V(D)J recombination 2) the imprecise joining of gene segments with nucleotide deletions and insertions at the segment junction sites and 3) the pairing of two unique somatically rearranged TCR chains. Usually, repertoire diversity is evaluated by analyzing the number of unique CDR3 motifs in a TCR repertoire. Estimates on the potential diversity, often referred to as the *theoretical repertoire* size, that can be generated through the above mechanisms vary substantially, ranging from 10^{15} [119] to 10^{61} [120]. The variance in these estimates depends largely on the number of nucleotide insertions and deletions that the underlying mathematical model accounts for. More recent estimates showed that

in humans, the CDR3β motifs alone exhibit a potential diversity of 10^{14} and the number of inserted nucleotides at the junction regions can be substantially higher than six as assumed by previous models [121]. Diversity calculations are further complicated by the fact that each unique CDR3 motif can potentially be generated through multiple different recombination events. These convergent recombination events are thought to be one of the main reasons for the emergence of public TCRs that are shared across different individuals at high frequency [122].

In any case the theoretical repertoire size is several orders of magnitude larger than the total number of T cells present in mice ($2x10^8$ [123]) and humans ($1x10^{12}$ [124]). For this reason, the *realized repertoire* represents a small fraction of the theoretical repertoire, especially in young individuals in which thymic output still provides a constant supply of new naïve T cells. In humans, the realized repertoire was estimated to consist of $1x10^6$ unique TCRβ chains that each pair with an average of 25 TCRα chains [124]. Consequently, the lower bound estimate for total diversity in the realized naïve human repertoire is $2.5x10^7$ unique TCRs. Interestingly, a PCR cDNA amplification-based assessment of the realized diversity in murine naïve TCR repertoires has provided evidence that despite the very different numbers of total T cells, TCR repertoire diversity is remarkably similar in mice and humans. According to a study published by *Casrouge et al.* [125] the αβ-TCR diversity in mice is approximately $2x10^6$. A possible explanation for this phenomenon was already postulated in 1987 by the definition of a functional unit termed the *protecton* [126]. The basic idea behind this concept is that there is a minimal repertoire size required for effective defense against the broad range of pathogenic threats and this minimal repertoire size can be defined as a unit that exists at different copy numbers and scales with the body size of an individual. More recent mathematical modeling of the required repertoire size for effective protection against a wide range of pathogens indeed provides evidence that the minimal diversity does not need to be much larger than initially hypothesized in the context of the protecton [127]. The same study also provides several potential explanations for the existence of a massive theoretical repertoire in the context of a relatively small protecton. Firstly, only about 5% of generated TCRs are selected during thymic selection [91] which significantly reduces the size of the realized repertoire. Secondly, TCRs are MHC-restricted and therefore must recognized

antigens presented by MHCs generated from thousands of different MHC alleles across populations.

A critical limitation of all the above size estimates of TCR diversity is their lack of information on the pairing dynamics of TCRα and TCRβ chains. Total diversity is commonly extrapolated based on bulk sequencing of single chains. The combination of limited throughput and/or immense costs per experiment makes the single-cell evaluation of entire repertoires unfeasible. The development of CITR-seq as a new method for high-throughput single-cell TCR sequencing as part of this PhD project can overcome many of the above limitations and significantly improve our ability to estimate TCR repertoire diversity.

Methods for TCR sequencing library preparation

In the past decades, a broad range of techniques have been developed for quantitative and qualitative TCR diversity analysis. Pioneering studies from the pre-high-throughput sequencing era utilized antibodies targeting specific TCR variants in combination with flow cytometry to gain first quantitative insights into TCR diversity [128]. PCA-based amplification of the CDR3 region of TCRs was applied in a technique called CDR3 spectratyping, in which the amplified DNA fragment-lengths were compared to quantify the frequency of TCR variants [129]. The development of (high-throughput) DNA sequencing provided access to evaluations of TCR diversity at the nucleotide level. Here, I will now focus on the different methods for preparation of TCR sequencing libraries.

The first decision to make when choosing a library preparation approach is whether to use DNA or RNA as input material. DNA is generally less sensitive to degradation and can be extracted from low-quality, ancient or formalin-fixed samples. However, due to the presence of intronic sequences, DNA fragments covering large portions of the V(D)J gene segment region are significantly longer than in RNA-based approaches and require additional DNA fragmentation to be suitable for short-read sequencing. The most limiting factor of DNA-based approaches is the presence of only two copies of the targeted TCRα and TCRβ loci. In contrast to that, mRNA transcripts exist at very high copy numbers for both TCR chains in each T cell [130]. Consequently, reverse transcription primers used for cDNA generation from TCR mRNA have significantly more targets, yet this also

complicates the quantification of TCR variants since each TCR can be expressed at different levels across T cells. Individual mRNA transcripts can be identified by the use of UMIs which enable the quantification of captured transcripts per unique TCR and can therefore be used to normalize for expression differences [131]. Especially in large datasets, in which the chance of capturing highly similar TCRs on both the V(D)J gene segment or CDR3 sequence level, UMIs can be used to distinguish PCR or sequencing errors from low-frequency TCR variants. Further, quantifying TCRs on the transcript level rather than the sequencing read level is more precise since it is less affected by biases introduced through variance in amplification efficiency of particular TCR variants [132]. Regardless of the choice of input material, TCR cDNA or gDNA needs to be amplified to generate sequencing libraries. A common strategy for this amplification is multiplexed PCR using forwards primers that are complementary to Vα and/or Vβ gene segments and reverse primers that target Jα/Jβ gene segment sequences (for gDNA-based approaches [133, 134]) or the TCRs constant region (for cDNA-based approaches [135, 136]). The design of primers that target those specific TCR regions requires previous knowledge of the underlying DNA or RNA sequence, which might not always be available especially in the case of non-model organisms. During multiplexed PCR, each individual primer can exhibit very different amplification efficiencies, even for primers with matched annealing temperatures. Extensive optimizations are required to adjust the individual primer concentrations in the final primer pool to compensate for those amplification efficiency biases [137]. To some extent, these biases can be addressed by grouping sequencing reads at the transcript level through the usage of UMIs, as described above [138]. A critical limitation of multiplexed PCR is that the usage of a distinct primer pools prevents *de novo* identification of unannotated TCR gene segments. This is not the case for approaches that are based on rapid amplification of 5’ complementary ends (5’ RACE) [139]. These RNA-based methods can be used to recover full-length TCR transcripts without previous knowledge of any TCR sequences [140, 141]. They build on template-switching which is facilitated by the addition of non-template nucleotides (cytosine in most cases) by Moloney murine leukemia virus (MMLV) reverse transcriptases [142]. After completing reverse transcription of mRNA, a template-switch DNA oligo can anneal to these non-template cytosines and the MMLV reverse transcriptase can use the oligo as new

template (by switching templates) to further extend 5' cDNA ends. This introduces a common sequence to all cDNA 5' ends that can be used as primer annealing site for subsequent cDNA amplification. The efficiency of template-switching reactions is generally low and depends on the precisely fine-tuned reverse transcription conditions with some inherent biases [143]. Collectively, the choice of multiplexed PCR or 5' RACE based methods depends on the specific type and quality of input material, as well as the required throughput and sensitivity for the data analysis [144].

All the above strategies can be modified to be used in single-cell TCR sequencing protocols. As highlighted before, single-cell sequencing of TCRs is necessary to pair TCRα and TCRβ chains expressed in the same original T cell. Since both TCR chains collectively define the antigen specificity of a given T cell, paired αβ-TCR information is crucial to identify target antigens of TCRs. It has been shown that different computational methods that aim to predict the target antigens of TCRs, perform significantly better when supplied with paired TCR data [145]. A common requirement for all single-cell TCR sequencing methods is the assignment of unique molecular barcodes to all TCR sequences originated from the same cell. This becomes increasingly more difficult with scaling numbers of T cells in an experiment. Thus, in comparison to bulk sequencing approaches, most single-cell methods are limited to 10^3 – 10^4 T cells in each experiment (reviewed here [146]). Early single-cell TCR sequencing methods separated individual T cells in multi-well plates using fluorescence-activated cell sorting (FACS) [147]. The physical separation of T cells into different wells effectively prevents cross-contamination leading to high confidence of αβ-TCR pairing but is very inefficient for high-throughput analysis and cost-intensive because all molecular reactions must be performed in hundreds of individual reactions. A major improvement of throughput was gained through the development of microfluidic systems for compartmentalized amplification reactions [148]. *Turchaninova et al.* modified this approach to first encapsulate single T cells in aqueous droplets in an oil emulsion and subsequently perform reverse transcription with barcoded primers within each droplet [149]. This way, all transcripts of a single cell captured in each droplet received a unique barcode that was used to pair TCRα and TCRβ chains. A common limitation of all microfluidic methods is the requirement for specialized instruments that perform the delicate cell encapsulation

process which is critical to the success of an experiment. Commercialized versions of microfluidic systems are now available and have become the dominant method of choice for most single-cell applications in present day studies (transcriptomic assays, epigenomic assays and TCR sequencing). Companies like 10x Genomics have developed straight-forward simple protocols for these “omics” applications, including TCR sequencing, that made them accessible to a broad range of laboratories, however they often come at immense cost. Commercial microfluidic devices and library preparation kits cost tens of thousands of dollars and therefore can quickly become unfeasible for many laboratories, despite the theoretical option of generating sequencing libraries for $>10^5$ cells.

The most recent expansion of throughput for single-cell applications is based on strategies involving molecular barcoding by combinatorial indexing of single cells. Initially designed for single-cell chromatin-accessibility analysis [150, 151], these methods are now also available for RNA sequencing [152] or even multiomic approaches capturing both modalities [153]. In those assays the cell itself functions as a reaction compartment for each molecular reaction. Molecular barcoding of single cells is achieved by a multi-step split and pool barcode ligation procedure in individual wells of multi-well plates. The combinatorial power of sequential addition of barcodes ensures that every cell’s “path” through the different ligation reactions results in a unique cellular barcode added to all gDNA/cDNA molecules of a cell. A common feature of combinatorial indexing-based methods is that they are extremely scalable and cost-efficient because all individual reactions are performed in bulk. With potentially more than one million cells that can be processed in a single experiment, the associated sequencing cost rather than the experimental throughput has become the limiting factor for single-cell experiments.

At the beginning of my PhD, we identified the potential of combinatorial indexing of single cells when applied in the context of TCR sequencing which led to the development of CITR-seq. Compared to whole-transcriptome or genome-wide chromatin accessibility methods, CITR-seq specifically targets just two transcripts (TCRα and TCRβ) which significantly reduces the required sequencing power per cell. To set this into perspective, in the CITR-seq data presented here, confident assignment of TCRα and TCRβ chains based on the presence of multiple transcripts per cell was possibly at a sequencing depth

of just 100 reads per T cell. In contrast to that, due to the expression of thousands of gene in each cell, whole transcriptome single-cell sequencing usually requires >20.000 reads per cell to capture a meaningful fraction of the expressed genes. On the other hand, based on the immense barcoding space generated by combinatorial indexing, cellular throughput is still in the order of hundreds of thousands of T cells in each CITR-seq experiment. In the course of the development of CITR-seq we evaluated the different library preparation approaches outlined above and integrated them into a combinatorial indexing framework. The final version of CITR-seq can be categorized as RNA-based approach that incorporates UMIs for transcript quantification. It further utilizes multiplexed PCR for cDNA amplification and acquires single-cell resolution through combinatorial indexing of individual T cells. A detailed description of the CITR-seq library preparation workflow is provided in the attached manuscript in chapter 2.

Methods for TCR repertoire analysis and comparison

In comparison to whole transcriptome sequencing data, the analysis of TCR sequencing data, often only containing information derived from just two genes, might appear to be much more simple. In fact, many studies that simultaneously profile the TCR repertoire alongside the transcriptional profile of T cells first generate whole transcriptome libraries and then specifically extract TCR-related reads from those libraries [154-157]. Despite encoding for just two genes, transcript diversity in a TCR dataset vastly outnumbers the diversity of protein-coding genes in mice and humans, even when isoforms are considered [158, 159]. This immense diversity is key to many of the challenges associated with TCR repertoire analysis. For example, the principle of exhaustive sequencing can hardly be applied in the context of evaluating repertoire completeness. Increasing the sequencing depth to a point at which additional sequencing reads do not yield previously unobserved transcripts is usually indicative of the sensitivity of underlying library preparation method and to some extent the completeness of the transcriptome. It has been shown that in two TCR libraries generated from a single human peripheral blood sample, about 75% of CDR3β sequences were unique to each library at saturating sequencing depth [160]. Especially in young individuals, in which the naïve T cell compartment is constantly replenished by recent thymic emigrants, TCR sequencing

libraries therefore represent a momentary fraction of the complete repertoire. Here, one can draw the first connection to ecological studies of biodiversity in distinct habitats. Famously known as the founder of the *unseen species problem* in the 1940s, Alexander Steven Corbet collected Butterflies in a distinct habitat in British Malaya for two years and wondered how many more he would identify after two years of additional collection [161]. Similarly, the total number of TCRs (species) in the complete repertoire (habitat) is unknown and expansion of the number of sampled T cells is likely to increase the number of identified unique TCRs. For this reason, different diversity estimators established in ecology are commonly used to estimate TCR diversity. The respective indices can broadly be classified into measurements of α- and β-diversity (not related to the two TCR chains) established by Robert Whittaker in 1960 [162]. α-diversity indices can be used to describe the diversity of TCRs within a repertoire, while β-diversity indices evaluate the diversity and/or overlaps across different TCR repertoires (e.g., of different individuals or repertoire diversity before and after infection). Several indices exist for estimating both diversity types (reviewed here [163, 164]). In the present study we used a normalized version of the Shannon diversity index (nSDI) [165], which takes into account both the relative abundance and the richness (e.g. total number of CDR3s or V(D)J gene segments depending on the level at which diversity is evaluated) of TCRs, to evaluate diversity within a given repertoire. The nSDI reaches its maximum (nSDI = 1) in the case that all CDR3 sequences or V(D)J gene segment usage frequencies are equally distributed in the evaluated repertoire. A second index that was used in the attached manuscript is the Jaccard similarity index [166]. The Jaccard index describes the overlap between two samples by dividing the number of shared elements by the union size of both samples. In the context of repertoire overlaps it can be used to evaluate the degree of CDR3 motif sharing in repertoires of varying sample sizes. Critically the sharing evaluated by the Jaccard index is based on 100% identical CDR3 amino acid motifs, which does provide only a limited view of the potentially shared ability of two repertoires to recognize identical antigens. It has been shown that TCRs cluster in specificity groups consisting of similar but not necessarily identical CDR3 motifs that recognize pathogen-derived antigens [167]. Therefore, rather than comparing identical CDR3s, it would be preferable to compare similar CDR3 motifs across repertoires. Pairwise comparison of peptide similarity is

commonly evaluated using a Blocks Substitution Matrix (BLOSUM) [168]. These matrices that were originally developed to score the similarity of evolutionary divergent proteins have been adopted to estimate the distance between two CDR3 motifs, which is correlated to their likelihood of recognizing similar antigens [169]. However, pairwise sequence alignment quickly becomes unfeasible for large TCR datasets. The number of required pairwise comparison scales quadratically with the number of input sequences, therefore quickly exhausting the computational capacity of most systems. Several algorithms have been developed to overcome these limitations and most of them build on comparing kmers extracted from each input sequence which are comparably easier to handle [170, 171]. Similarly, in the attached manuscript, amino acid 4mers were used to evaluate repertoire similarities across the different mouse species.

With antigen specificity being the primary focus of TCR analysis, the CDR3 regions of both TCR chains are the sole focus of many studies. However, as discussed before, the CDR1 and CDR2 regions receive increasing attention because of their postulated role in TCR-MHC binding modulation [22, 23, 172]. Despite the fact that those sequences are germline-encoded, their identification from TCR sequencing data can be difficult due to the high sequence identity of particular (duplicated) V gene segments even when full-length TCR data is available [173]. D gene segments of the TCRβ chain are extremely short (Trbd1: 12 nt and Trbd2: 14 nt) and can deviate from the germline sequence in the majority of their mapped sequence due to nucleotide insertions and deletions. Further, these random insertions of nucleotides at the gene segment junctions are often difficult to distinguishing from sequencing or PCR errors. For this reason, commonly used alignment tools for RNA-seq data often perform poorly in TCR transcript alignment and specialized alignment tools have been developed [174-176]. In the present study we have used MiXCR [174] a Java-based software tool that uses a kaligner approach modified from *Liao et al.* [177] to map raw sequencing reads to a V(D)J gene segment reference.

With the advent of machine learning, the available tools have also been used to predict TCR epitopes from (paired) CDR3 sequences and/or V(D)J gene usage [178-180]. These tools aim to identify the cognate antigens of large sets of TCRs, whose antigen-specificity has not been experimentally validated. Critically, this analysis is complicated not only by the diversity of TCR repertoires but also by the diversity of MHC-haplotypes responsible

for antigen presentation. The available approaches can be broadly classified into supervised and unsupervised prediction models. Supervised models are supplied with experimentally validated TCR-antigen pairs and base their prediction on these training datasets [181]. In contrast to that unsupervised models, unsupervised models apply TCR distance-based prediction using the algorithms described before [169, 171]. Significant performance differences have been seen across those prediction tools and their ability to infer the antigen specificity of previously unseen TCR motifs is limited (reviewed here [182]). We expect that high-throughput methods like CITR-seq can make important contribution to the training of machine learning models by supplying an extensive wealth of experimentally validated TCR pairs.

Mouse models in studies of the adaptive immune system diversity

The majority of studies that established the pioneering concepts and led to many breakthrough discoveries in the field of (adaptive) immunity were and are still conducted using established mouse models. Because of their easy husbandry, short generation time and relatively recent latest common ancestor to humans (about 85 million years ago [183]), mice are by far the most widely used mammalian-model system in biomedical research (e.g., almost 75% of all laboratory animals in 2022 in Germany [184]). Most of today's laboratory mice are derived from common inbred strains that were first established about 100 years ago (e.g., C57BL/6 in 1920s by C.C. Little) to reduce the impact of genetic variance on research findings from different mouse studies. While this has significantly improved study reproducibility, it creates a paradox in the context of studying the natural diversity of adaptive immune systems. For example, while outbred populations of mice [185] and humans [17] display remarkable MHC-haplotype diversity, all inbred C57BL/6 mice share the identical MHC-haplotype H-2^b. Consequently, the presented immunopeptidome as well as thymic selection of T cells in laboratory mice might not be representative of the diversity and dynamics observed in the underlying processes in outbred populations [20, 186]. The literature on TCR diversity in outbred mice compared to laboratory strains is extremely sparse [187, 188]. These studies focus on establishing the orthology of V(D)J gene segments in different murine sub-species. In the context of this PhD project, we showed frequent copy number variations (CNVs) and nucleotide

polymorphisms in Vα gene segments, even in closely related mouse species. In agreement with this, V gene segments with missing murine orthologs have also been identified in outbred bank vole populations [189]. The same study also reported remarkable inter-individual V(D)J gene segment usage biases, which are likely caused by the diverse MHC-haplotypes in the studied bank vole population. In conclusion, the severely restricted genetic diversity in inbred laboratory mice is likely to have a significant impact on population-scale TCR diversity studies. To date, this topic has gained very little attention and is therefore poorly understood.

Apart from limited genetic variance in laboratory mice, their husbandry in specialized facilities creates another paradox for evaluating TCR diversity. The adaptive immune system has evolved to recognize an immensely diverse range of pathogenic challenges and establish long-term immunity against those threats following the initial encounter in the hosts environment. Yet, in order to minimize the impact of environmental noise, laboratory mice are housed in “clean” and sometimes even pathogen-free facilities. Accordingly, large differences in various measurements of immune functions have been identified when comparing “wild” mice to laboratory strains [190]. As a general trend, wild-caught mice showed greater variance in most measurements of immune function. Immune challenges using sheep red blood cells in wild-caught and laboratory mice showed significantly more effective clearance of these cells in wild-caught mice, likely because their immune systems were primed from previous antigen exposures [191]. Differences in key immunological processes are even more pronounced when comparing laboratory mice to humans (reviewed here [192, 193]). These differences, along with the common failure to translate immunological research findings from mice to humans, has led to repeated questions about whether studies of the murine immune system are representative of human immunology. Interestingly, co-housing laboratory mice with pathogen-exposed pet-store mice induces changes in response to infection, T cell differentiation and general immune cell gene expression patterns, that more closely represent patterns observed in humans [194]. For example, after co-housing, the fraction of naïve $CD8^+$ T cells relative to the fraction of effector $CD8^+$ T cells was significantly reduced and more similar to the human fractions as a consequence of persistent pathogen-exposure and acquisition of a growing memory T cell compartment. Although

not investigated in this study, it is very likely that these shifts in T cell populations would also lead to significant changes in TCR repertoire diversity. In summary, there now is a general consensus that “naturalizing” laboratory mice [195] might alter some immunological functions to a state that is more representative of outbred populations of mice and humans.

In contrast to that, far less attention has been paid to acknowledging the impact of limited genetic diversity in the adaptive immune system of laboratory mice. This gap in knowledge has been a substantial motivation for the presented PhD project. Studies on the generation of TCR diversity are either done in humans, exhibiting strong genetic variance especially in HLA-haplotypes, or alternatively, in a single inbred mouse line with almost no genetic variance across individuals. In both cases, the ability to analyze the impact of genetic variance on TCR repertoire selection and diversity is limited. In the manuscript presented in chapter two of this thesis, we used a collection of four evolutionary diverged inbred mouse species and their F1 hybrids, to investigate the dynamics of TCR repertoire generation in a distinct but much broader genetic context. The four respective inbred species that were originally derived from wild-caught mice are: PWD/PhJ (Jackson Laboratory strain ID: 004660, from now on PWD) an inbred strain of *Mus musculus musculus* caught in 1972 in the Czech Republic [196], CAST/EiJ (Jackson Laboratory strain ID: 000928, from now on CAST) and inbred strain of *Mus musculus castaneus* established in 1971 [197], SPRET/EiJ (Jackson Laboratory strain ID: 001146, from now on SPRET) an inbred strain of *Mus spretus* originally caught in Spain in 1978 [198] and the most commonly used laboratory mouse strain C57BL/6J (Jackson Laboratory strain ID: 000664, from now on BL6). Genomic studies in C57BL/6 have provided evidence that the largest fraction of its genome is derived from *Mus musculus domesticus*, with smaller introgressions from *Mus musculus musculus* and *Mus musculus castaneues* [199]. As indicated by their names, *Mus musculus domesticus*, *Mus musculus musculus* and *Mus musculus castaneues* are subspecies of the major *Mus musculus* lineage commonly referred to as the house mouse [200, 201]. Their classification as separate species represents an ongoing debate based on the presence of stable hybrid zones in wild populations of these mice [202, 203]. Critically, all the above inbred laboratory strains can form viable and, in some cases fertile offspring, making it possible

to investigate phenotypic differences of the parental lines in a common F1 hybrid genetic background. Today, thanks to their fully sequenced genomes [204], these alternative laboratory mouse strains, provide an exceptional resource for a broad range of studies of speciation, adaptation and the genetic basis of complex traits (reviewed here [205, 206]). Several major phenotypic (immunological) differences have been identified across these inbred strains. Perhaps the most interesting in the context of T cell biology are major differences in *Fas* death receptor expression which is critical for T cell activation, proliferation and apoptosis [207], and hyperresponsiveness to high doses of tumor necrosis factor, a crucial pleiotropic proinflammatory cytokine [208]. These immune related phenotypes were further investigated in collaborative crosses of the respective inbred lines revealing major differences in the frequencies of specific T cell populations [209]. To the best of our knowledge, inbred mouse lines have never been used to reveal the impact of genetic factors on the generation of diverse TCR repertoires. Comparing their unique sets of V(D)J gene segments in terms of usage and selection has immense potential to expand our knowledge of TCR repertoire generation. This is especially true for TCR analysis in F1 hybrids, in which the different sets of parental V(D)J gene segments are subject to thymic selection in defined heterozygous MHC-haplotypes.

Multi-omic analysis of gene regulation

The versatile combinatorial barcoding system applied in CITR-seq has also been modified to be used in the much broader context of studying the regulation of gene expression. The development of easySHARE-seq, presented in chapter 3, allows for the simultaneous measurement of gene expression and chromatin accessibility at single-cell resolution. Here, I will now briefly introduce the advantages of utilizing such a multi-modal approach for analyzing the regulation of gene expression and outline the potential of its application in T cell biology.

Approaches that aim to characterize the transcriptome of a single cell have now been available for roughly 15 years [210]. Until then, bulk sequencing of a heterozygous collection of cells from various tissues only provided information about the average expression of a gene across all cell types in the respective tissue or sample. It has been known for a long time, that gene expression variance in different cells is one of the main

sources of phenotypic variance [211, 212], especially in heterogenous tissues such as tumors [213]. Even with single-cell technologies at hand, that provide the power to cluster individual cells by cell types based on their transcriptional profiles, many fundamental questions regarding phenotypic variance remained unanswered. For example, while whole transcriptome analysis of single cells can reveal gene expression differences, the causes of variance in gene regulation can hardly be inferred from transcriptome data alone. Nonetheless, a plethora of those (non-coding) genome-transcriptome associations that mediate gene expression differences had been known based on individual examples (e.g., [214] and reviewed here [215, 216]). Consequently, the simultaneous analysis of additional modalities, such as epigenomics or proteomics, is required to overcome these limitations. Recently, chromatin accessibility has gained special attention because chromatin states of either hetero- or euchromatin have significant implications for transcriptional activity [217, 218]. Today, the state-of-the-art approach for (single-cell) chromatin accessibility studies is the **a**ssay for **t**ransposase-**a**ccessible **c**hromatin (ATAC-seq) [219]. The field of single-cell biology has rapidly accelerated and various combinations of sequencing methods have been integrated to into “multiomic” approaches (reviewed here [220]). Arguably the most widely used multiomic approach is the combination of ATAC- and RNA-seq and custom protocols as well as commercial platforms (e.g. by 10x Genomics) have been developed. For instance, SHARE-seq has been used to show that chromatin accessibility changes precede changes in gene expression during murine hair follicle differentiation [221]. In the same study a computational strategy was developed to reveal potential *cis*-mediated gene regulation based on cell-specific co-variance of distal ATAC-seq peaks and gene expression. This enables the systematic genome-wide prediction of regulatory elements driving gene expression variance that can subsequently be validated by functional assays at fine-scale. EasySHARE-seq represents a refined version of the original protocol, enhancing flexibility and RNA-seq sensitivity. The increase in flexibility mostly relates to the implementation of the combinatorial barcoding system that was also utilized in CITR-seq. By using this system, it is now possible to multiplex easySHARE-seq libraries with other sequencing libraries and jointly sequence them using Illumina sequencing devices with standard index cycle length configurations.

Although not done in the scope of this PhD, the application of easySHARE-seq to various T cell populations in health and disease has enormous potential. For example, T cell lineage commitment in the thymus is characterized by the orchestrated expression of multiple transcription factors (reviewed here [222]). Disentangling those complex interactions could be achieved by tracking their expression alongside changes in chromatin accessibility at their potential target sites. Further, as outlined earlier in this introduction, V(D)J recombination order is guided by precisely timed chromatin remodeling, leading to accessibility changes of RSS sequences of individual gene segments. Multiomic assessment of the expression of genes of the core recombination machinery as well as changes in V(D)J gene segment chromatin accessibility is likely to provide further insight into the fine-scale underlying mechanisms.

Objectives

Analysis of the TCR repertoire can provide crucial insights into the past, present and future immune responses of an individual following the exposure to pathogens or other malignancies. Technical and financial barriers of current TCR sequencing approaches are still limiting our ability to analyze the magnitude and diversity of TCR repertoires, as well as the relative contribution of stochastic and genetic factors in the generative process. For this thesis, I, together with my colleagues, have developed a new experimental protocol for the large-scale analysis of single-cell TCR repertoires. This platform allowed us to address some long-standing questions in TCR biology. To what extent is an individual's TCR repertoire shaped by genetic factors? Could such genetic factors provide evidence for the co-evolution of TCRs and MHCs? What is the mechanistic basis for the high frequency of public TCR motifs observed between individuals? In my PhD project, I have utilized panels of wild-caught inbred mouse species and their F1 hybrids and took advantage of their distinct genetic backgrounds to address these questions.

For chapter one, we have conducted a comprehensive cross-species comparison of TCR V(D)J gene segment loci in four murine inbred lab-strains. We highlight that gDNA-based gene segment annotations are often incomplete because of gaps in the respective genome assemblies caused by the complexity of the underlying loci. Further we highlight remarkable diversity in the TCRα variable gene segments across murine sub-species. For instance, we report a recent major locus contraction in *Mus musculus castaneus* which lead to the loss of 74 Trav gene segments. This effort also aimed to generate a detailed sequence library of V, D and J gene segments, which is required for the fine-scale mapping of sequencing reads originated from TCR repertoires of the respective mice.

In chapter two, comprising the main work of my PhD, I first elaborate on the development of CITR-seq, a flexible, high-throughput and single-cell TCR sequencing approach that allowed us to generate paired αβ-TCR repertoires from 32 individual mice. Collectively,

their repertoires consist of more than 5 million receptors and therefore likely represents the largest dataset of confidently paired TCRs analyzed in a single study to data. The generated species-specific V(D)J gene segment references allowed us to investigate the differences of their usage frequencies across the different mouse species. This revealed that intra-species usage frequencies are remarkably conserved. We then used in-frame and out-of-frame TCR receptor sequences to specifically evaluate the impact of thymic selection on the shaping of the TCR repertoire. Finally, the joint effects of species-specific generation and selection of TCRs were evaluated at the level of CDR3 sequence diversity and in the context of CDR3 motif sharing across different mice.

Chapter three summarizes the development of easySHARE-seq, a multi-omic protocol that can be used to simultaneously assay the transcriptome and chromatin profile of single cells. I contributed to this project by participating in the development of the utilized single-cell barcoding system, that is similar to the barcoding approach applied in CITR-seq. In future studies, this method offers great potential to expand the analysis conducted in chapter two by also integrating whole transcriptome and chromatin accessibility profiles in the analysis of TCR repertoires.

In the discussion section, I will recapitulate the findings from all chapters, integrating them with recent research and current debates in TCR biology. Furthermore, I will provide an outlook on future research directions for TCR analysis, highlighting their potential in current and future medical applications to combat diseases.

Chapter 1: Distinct evolution at TCRα and TCRβ loci in the genus *Mus*

Moritz Peters[1*], Volker Soltys[1], Dingwen Su[1], Yingguang Frank Chan[1,2*]

1. Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany
2. University of Groningen, Groningen Institute for Evolutionary Life Sciences, 9747 AG Groningen, Netherlands

*Corresponding authors: moritz.peters@tuebingen.mpg.de, frank.chan@rug.nl

Status in submission process: Advanced manuscript

Abstract

T cells recognize an immense spectrum of pathogens to initiate immune responses by means of a large repertoire of T cell receptors (TCRs) that arise from somatic rearrangements of *variable*, *diversity* and *joining* gene segments at the TCR loci. These gene segments have emerged from a limited number of ancestral genes through a series of gene duplication events, resulting in a greatly variable number of such genes across different species. Apart from the complete V(D)J gene annotations in the human and mouse reference assemblies, little is known about the structure of TCR loci in other species.

Here, we performed a comprehensive comparison of the TCRα and TCRβ gene segment clusters in mice and three of its closely related sister species. We show that the TCRα *variable* gene cluster is frequently rearranged, leading to deletions and sequence inversions in this region. The resulting complexity of TCR loci severely complicates the assembly of these loci and the annotation of gene segments. By jointly utilizing genomic and transcriptomic data, we show that in *Mus musculus castaneus* the variable gene cluster at the α locus has undergone a recent major locus contraction, leading to the loss of 74 *variable* gene segments. Additionally, we validated the expression of functional variable genes, including atypical ones with inverted orientation relative to other such segments. Disentangling the fine-scale structure of TCR loci in different species can provide valuable insights in the evolution and diversity of TCR repertoires.

Introduction

T cells are the principal cell type underlying adaptive immunity and perform the remarkable task of distinguishing self from foreign to decide whether or not to initiate an immune response. This pivotal decision depends solely on the recognition of antigens presented by major histocompatibility complexes (MHCs) by the heterodimeric TCR. To cover the enormous space of potential pathogenic antigens, a vast diversity of specific TCRs is required. Most T cells express a unique TCR consisting of an α- and a β-chain that arise from somatic rearrangements in a process called V(D)J recombination [1]. Estimates of the diversity generated by this recombination process vary substantially and range from 10^{15} [2] to 10^{61} [3] depending on the mathematical model and the evaluated species. In any case, these theoretical estimates of diversity are several orders of magnitude larger than the observed diversity in any individual (e.g., $2x10^{8}$ in mice [4] and $1x10^{12}$ in humans [5]), due to the significantly lower number of total T cells and diversity reduction by selection of specific TCRs during T cell maturation.

The building blocks of TCRs are the variable (V), diversity (D, exclusive to TCRβ) and joining (J) gene segments that are subject to somatic rearrangements by V(D)J recombination. The underlying process requires the precise execution of an ordered series of DNA double-strand breaks that is facilitated by the *Rag1*/*Rag2* recombinase complex [6]. The respective double-strand breaks are repaired by non-homologous end joining (NHEJ) [7] during which random insertions and deletions of nucleotides can occur, which further increases TCR diversity [8]. Recombination signal sequences (RSS) that are interspersed between V(D)J gene segments are targeted by the *Rag1*/*Rag2* complex to initiate recombination. These conserved sequences consist of a heptamer sequence, a spacer sequence with a conserved length of either 12 or 23 base pairs and a conserved nonamer sequence [9]. The so called 12/23 rule ensures that V(D)J recombination results in the fusion of a single V to (D) to J segment [10]. The distinctive sequence features of RSS's are reminiscent of sequences of transposable elements [1]. It is therefore likely that an ancestral version of a TCR gene segment has been invaded by a transposon and subsequently the split gene had to be recombined to encode a functional protein. This hypothesis is supported by the presence of several TCR and BCR related genes in lower chordates that represent potential targets of the initial transposon invasion [11]. V and J

gene segment sequences have been categorized into complementarity determining regions (CDR) and framework regions (FR) based on the position of highly conserved amino acid residues in their coding sequence (e.g., cysteine at position 23 and 104 of V genes [12]). The germline-encoded CDR1 and CDR2 regions in the coding sequence of V genes have been shown to modulate TCR-MHC binding affinity [13], while the CDR3 region that comprises the highly diverse junctional region of V(D)J gene segments mainly determines the antigen specificity of the TCR [14, 15].

The distinctive features found in the coding sequence of V(D)J gene segments as well as in their sequence vicinity (e.g., RSS) have allowed their identification from genomic sequences even in the absence of detailed gene annotations [16]. These approaches have revealed that the number of functional V(D)J gene segments varies substantially across taxa and even between closely related species [17-20]. V gene segments are often grouped into families with one to twelve members depending on their sequence similarity of ~70-100% [21]. TCRβ exclusive usage D gene segments is highly conserved as well as an expansion of the number of J gene segments in the TCRα chain [22]. Similarly, the number of variable gene segments also varies among immunoglobulin heavy chains across different mammals [23]. Successful inference of functional gene segments, however, depends largely on the quality of the respective genome assembly, with complex loci like TCR often representing the worst assembled loci in non-model organisms. This has been emphasized by a recent study that identified V and J genes in the bank vole based on transcriptomic data and identified several additional genes that had not been identified from the genomic sequence [24]. In the same study, most of the identified TCR V and J gene segments were shown to have clear murine orthologs except for three of the identified V genes. In general, significantly less is known about J and D gene segment gene cluster variance, likely because of their relatively short sequence with fewer distinctive features, making it challenging to identify those genes in different genomes. Nonetheless, because both J and D genes contribute mainly to the antigen specificity rather than TCR-MHC binding, the evolution and diversification of their respective loci is particularly interesting in the context of host-pathogen co-evolution.

Locus expansion and contractions of TCR gene segment regions is consistent with the birth-and-death hypothesis of multigene families [25, 26]. This hypothesis is derived from

the observation that sequence homogeneity between members of a gene segment cluster within a species is not necessarily higher than to gene segments of a different species [27-29]. It provides the mechanistic basis to explain the evolution of divergent gene segment families, including high frequencies of non-functional and pseudogenized genes following gene duplications and release of functional constraint due to redundancy. While initially evaluated in immunoglobulin and MHC families, subsequent comparative studies of TCR V gene segment families confirmed that sequence identity of homologous families in mice and human exhibit higher similarity than observed between intra-individuals gene families [30]. Later, this view was expanded by showing that divergent Vβ gene segments have been maintained in murine and human genomes for more than 100 million years, strongly indicating that the initial gene duplication events are ancient and predate the split between human and mice [31]. In this context it is important to highlight, that while the diversification of immunoglobulin receptor and TCR loci appears to be driven by similar mechanisms, MHC restriction of TCRs might impose that duplicated gene segments maintain the ability to bind to MHCs to stay functional. In contrast, immunoglobulin gene segments can diversify without such inherent requirements. There now is evidence that four ancestral Vβ and five ancestral Vα gene segments formed the original set of V genes at the root of all mammalian clades. These have since amplified and diverged to different extent in present day mammals [18]. In summary, the birth-and-death hypothesis therefore challenges the classical model of concerted evolution which states that multigene families emerge by inter-locus recombination alongside gene conversion so that all genes within a family cluster evolve in concert and homogenize over time [32].

The murine TCR Vα locus has been subject to one of the most drastic reported locus expansion events in which more than two-thirds of the central locus region has become amplified [21]. Strikingly, this duplication was estimated to have occurred just 4-8 million years ago but has received little attention so far. Today we have access to the high-quality genome assemblies of several common inbred murine strains as well as wild-derived sister species of the most common C57BL/6 laboratory mouse strain [33]. These wild-derived inbred strains share their latest common ancestor about 1-3 million years ago [34, 35] and therefore represent an excellent system to study the evolution of complex

traits (reviewed in [36]). Strikingly, the regions with the greatest sequence diversity within the assembled genomes of the various strains relative to the mouse reference genome (GRCm38/mm10) were found to be regions related to immune- or sensory-functions [37]. To date, most comparative studies of adaptive immunity in inbred strains or wild-caught mice are centered around quantifying immune cell populations or measuring differences in immune responses [38, 39]. In contrast, little is known about comparative genomics of TCR gene segment loci despite the fact that those are subject to frequent genomic rearrangements which likely cause significant differences in TCR diversity. Here, we provide a comprehensive comparison of murine TCR loci. By utilizing both, genomic and transcriptomic data, we highlight major rearrangements in the Trav locus of the wild-derived inbred mouse stain CAST/EiJ relative to the mm10 reference and thus emphasize the variability in these loci even in closely related species.

Results

The murine TCRα and TCRβ loci in the GRCm38/mm10-based reference

The genomic sequences of the TCR loci have been extensively studied in human and mouse. The gene annotations derived from these studies have been summarized in databases [40] which are now considered to contain all expressed V(D)J gene segments of both species. Here, we specifically focus on the murine TCR gene segments that are annotated in the IMGT database based on mouse reference genome assembly GRCm38/mm10 (from now on referred to as mm10 assembly). The TCR regions in this database are located in between genes referred to as “locus bornes” (French for milestone) which flank the TCR loci and display an evolutionary conserved gene order across taxa. These can therefore aid the localization of the respective loci. For example, the gene *Dad1*, a 3’ borne, marks the 3’ end of the TCRα loci cluster.

Murine TCR gene segments are found in clusters of varying numbers of gene segments and gene segments within a cluster are further grouped into families bases on their sequence homology and ancestry. The current reference TCRα loci consist of a total of 191 gene segments. These can be further divided into 130 Trav gene segments (including 20 pseudogenes, **Fig. 1A**), 60 Traj gene segments (including 12 pseudogenes, **Fig. 1B**) and a single constant region. All TRCα gene segments are located on chromosome 14

and collectively span about 1.8 Mbp (14C1, 26.94 cM – 27.70 cM). The majority of Trav gene segments consist of two exons with an average span of 556 bp. About two-thirds of the ancestral murine Trav cluster have been triplicated in a recent gene duplication event [21] and all triplicated genes were annotated with a “d” or “n” in their official names to indicate their origin in the ancestral locus configuration. Traj gene segments are all encoded by a single small exon with an average length of 59 bp. The antigen-specificity defining CDR3 region of the TCRα chain consists of the most 3’ bases of a V gene segment and the most 5’ bases of a J gene segment.

The TCRβ locus spans about 0.8 Mbp and is located on chromosome 6 (6B1, 18.93 cM – 19.71 cM). It consists of 35 Trbv genes (including 13 pseudogenes, **Fig. S1A**), 2 Trbd genes, 2 Trbc genes and 14 Trbj genes (including 2 pseudogenes, **Fig. S1B**). A unique feature of the TCRβ locus is the presence of an inverted V gene segment (*TRBV30* in human and *Trbv31* in mice) at the 3’ end of the locus. Both its position and orientation are conserved in all tetrapods [19]. Across different species the Dβ-Jβ-Cβ clusters are present at varying copy numbers (e.g. 2 in human and mice, 1 in chicken and 3 in swine, [41]). Due to the incorporation of D gene segments, the rearranged TCRβ transcript contains two junctional sites compared to the single junction in the rearranged TCRα transcript.

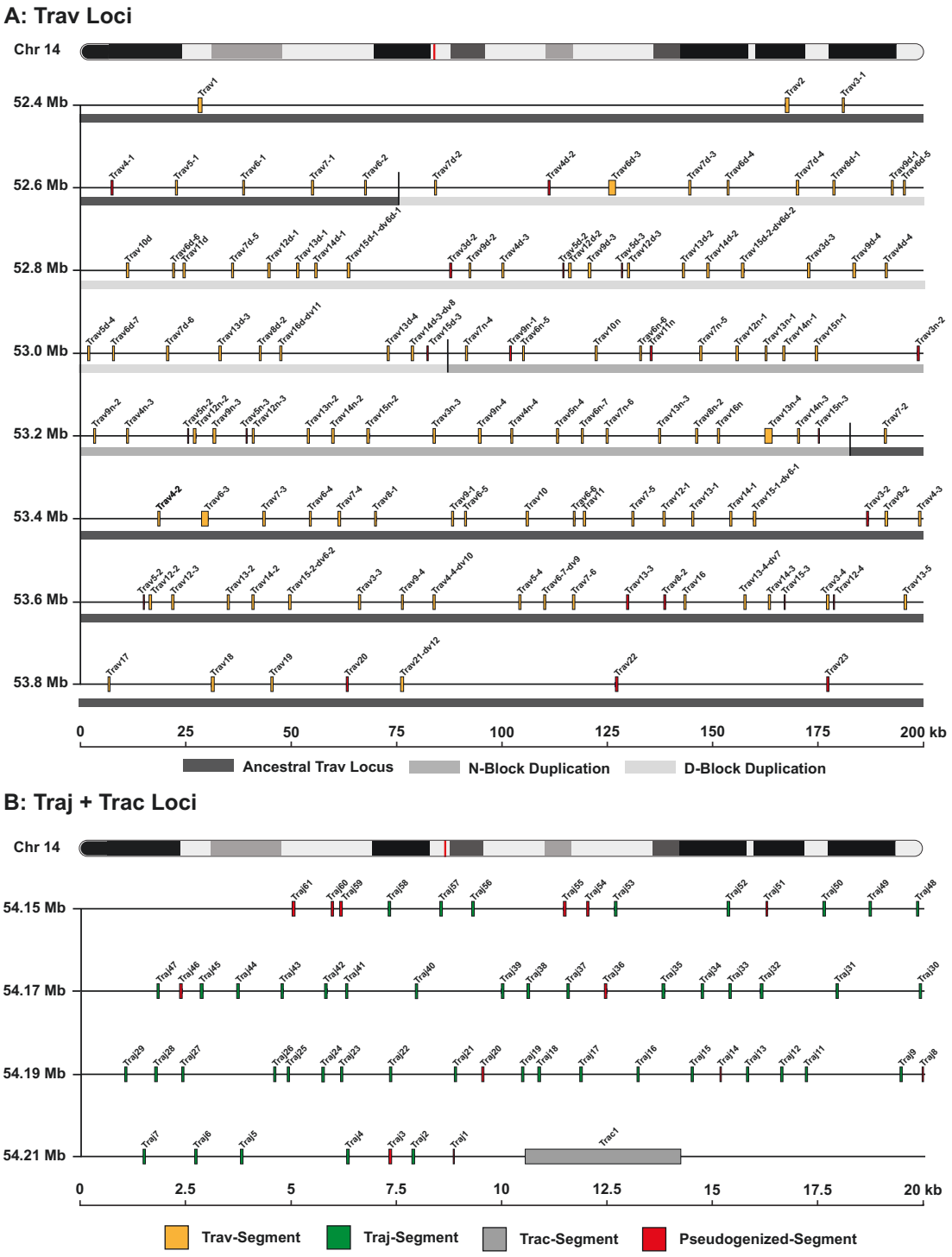


Figure 1: Genomic locations of V/J gene segments in the TCRα loci as annotated in the GRC38/mm10 based IMGT annotation. TCRα *variable* gene segments (130 segments, **A**) are clustered in a 1.6 Mbp genomic region on chromosome 14. A recent gene duplication event led to the triplication of roughly two thirds of the ancestral Trav loci (darkgrey) resulting in the d- (lightgrey) and n-block (middlegrey) Trav segments. Traj and Trac genes (60 Traj and 1 Trac; **B**) are clustered in an 80 kb window upstream of the Trav cluster. Gene segments annotated as pseudogenes are colored in red.

The evolution of murine TCR loci

To provide an overview of the TCRα and TCRβ gene segment clusters in other murine species we performed a pairwise sequence alignment of V, D and J gene segments alongside the constant regions in four different inbred mouse lines (129S1/SvImJ, PWK/PhJ, CAST/EiJ and SPRET/EiJ, from now on referred to as: 129, PWK, CAST and SPRET). For all four species, genome assemblies are made available by the Mouse Genome Project [33]. The dotplot of the local alignments of the Trav cluster confirmed the previously reported locus expansion by triplication of the central region of the cluster in 129, PWK and SPRET (**Fig. 2A**). Strikingly, this local sequence triplication was not observed in the CAST genome assembly and the entire cluster was contracted to a size of about 0.86 Mbp. Apart from this obvious size difference, we also observed local sequence inversions in the Trav cluster which were most frequent in the SPRET assembly relative to the mm10 assembly (**Fig. 2A bottom**). We frequently observed gaps in local assemblies in the genomes of all four mouse strains within the Trav clusters, which in part reflects the complexity of these loci. As a first approach to transfer the reference annotations, we performed a sequence liftover, complemented with a six-frame translation BLAT to identify the chromosomal locations of annotated mm10 V(D)J genes in the assemblies of the four murine inbred lab species (**Table 1**).

Table 1: Number of V(D)J gene segments (including pseudogenes) identified in the different inbred mouse strains.

SEGMENT/STRAIN	MM10	129	PWK	CAST	SPRET
TRAV	130	80	97	45 (58)	75
TRAJ	60	59	59	59	59
TRAC	1	1	1	1	1
TRBV	35	35	35	34	35
TRBJ	14	14	14	14	14
TRBD	2	2	2	2	2
TRBC	2	2	2	2	2

We did not observe any major locus rearrangement for Jα and Cα (**Fig. 2B**) as well as Vβ, Dβ, Jβ and Cβ clusters and the respective cluster sizes were highly similar among all strains (**Fig. 2C and 2D**). While the *Trbv31* inversion was found in all four strains, we did not observe any further inverted segments. In the following analysis, we now specifically

compare TCR loci of CAST to the mm10 reference to highlight the shortcomings of currently available V(D)J gene segment annotations for such complex loci.

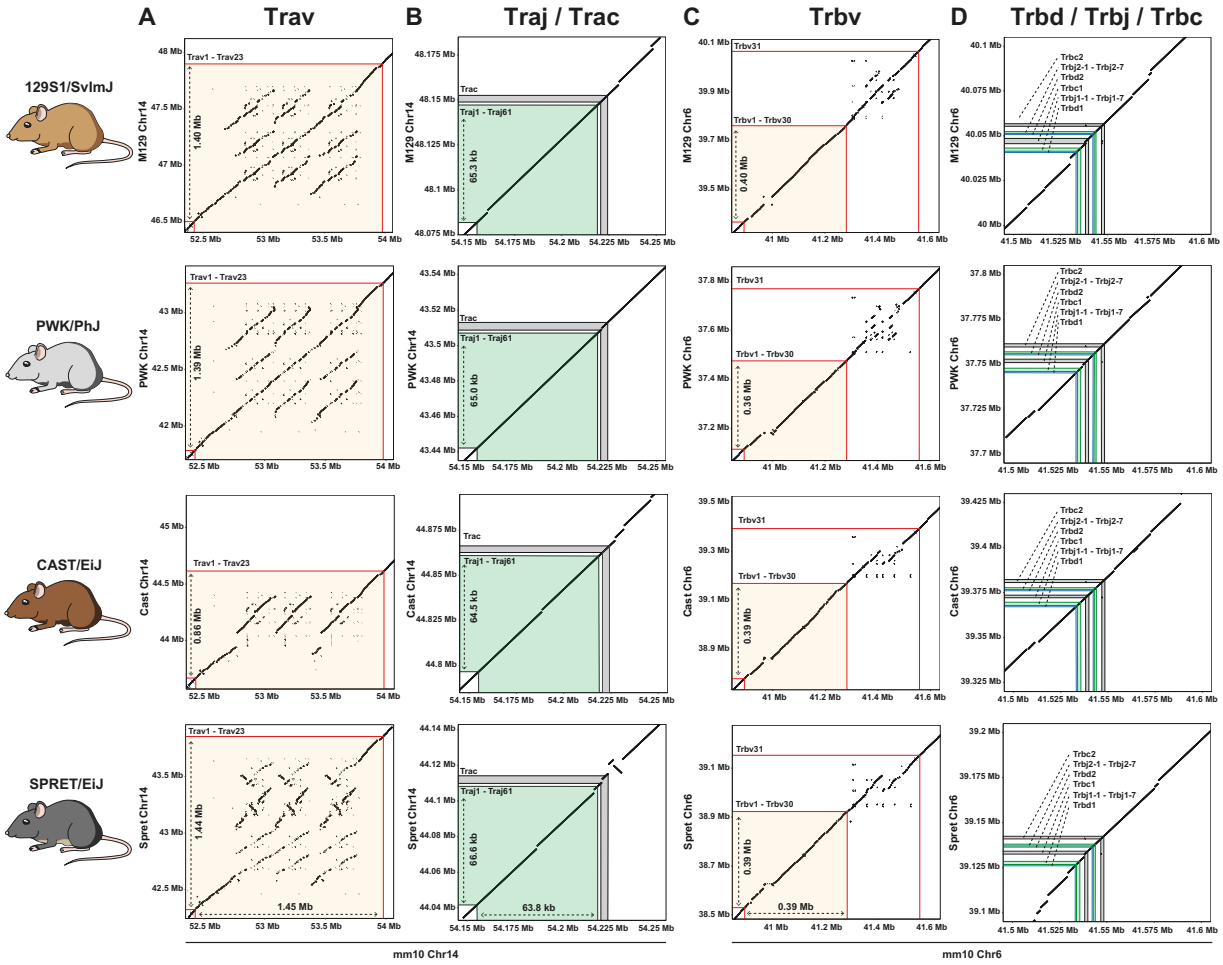


Figure 2: Pairwise alignment dotplots of the TCR genomic regions of four inbred laboratory mouse strains to the respective regions in the murine mm10 reference sequence. Shaded areas illustrate the genomic region ranging from the most 3' to the most 5' gene segments.

The TCRα and TCRβ locus in *Mus musculus castaneus*

The majority of currently available V gene prediction tools used for the *de novo* annotation of *variable* gene segments in non-model organisms identify candidates by sequence homology to known genes and/or identification of the highly conserved RSS sequences in the vicinity of gene segments [42, 43]. Inherently, these approaches depend on a gapless assembly of the underlying loci, which is often unavailable due to the high complexity V regions.

For CAST we identified a total of 79 full-length *variable* gene segments mapped to the CAST Trav (45) and Trbv (34) region (chromosome 14: 43.65 – 44.65 Mbp, **Fig. 3A** for Trav and chromosome 6: 38.75 – 39.40 Mbp, **Fig 3B** for Trbv). We showed that a large deletion led to the loss of a total of 74 Trav gene segments in CAST relative to the mm10 locus. The deleted Trav segments largely overlap with the triplicated segments in the recent murine Trav triplication event dated back to about 4 - 8 million years ago. Close inspection of Trav sequences alignments against its possible homologs revealed that some of the Trav gene segments in CAST exhibit greater similarity to the corresponding segment in the expanded D-block cluster than to the respective ancestral gene segment. For example, the CAST *Trav6(d)-4* gene showed 100% sequence identity with the mm10 *Trav6d-4* but only 97.5% sequence identity with ancestral *Trav6-4*. Pairwise sequence homology comparison in the remaining Trav gene segments revealed that the deletion junctions are likely located in between the CAST Trav gene segments *Trav7d-4* and *Trav8-1*. We therefore showed, that the “d-“ and “n-blocks” in the Trav locus were present in the ancestors to CAST and thus the present-day CAST Trav locus has undergone a secondary locus contraction, leading to the loss of the majority of the Trav segments in those blocks. All CAST Trav gene segments were subsequently annotated based on the gene segments showing the closest sequence homology in the mm10 reference. Taking into account the latest common ancestor [44] of 129 (*Mus musculus domesticus*), PWK (*Mus musculus musculus*) and CAST (*Mus musculus castaneus*), this contraction has likely happened less than 500,000 years ago. In addition, eleven Trav segments outside the major deletions could not be identified by lifting the genomic coordinates from mm10 to the CAST assembly. For 9 of those Trav segments (*Trav3-1*, *Trav13-1*, *Trav14-1*, *Trav12-2*, *Trav3-3*, *Trav13-3*, *Trav14-3*, *Trav3-4* and *Trav13-5*) we identified corresponding sequence fragments that were terminated by gaps in the CAST genome assembly. Interestingly, the identified *Trav3-1* fragment, consisting of just the first exon, showed an inverted sequence orientation (**Fig. 3B**) relative to the homologous sequence in mm10. The two remaining Trav gene segments, *Trav6d-3* and *Trav16*, were not found in the CAST Trav locus because of local sequence deletions (**Fig. 3C**).

Except for *Trbv9*, an ortholog of all 35 mm10 Trbv segments was successfully identified in the CAST Trbv locus. This was also true for all J gene segments across both loci (60 Traj and 14 Trbj gene segments; data not shown).

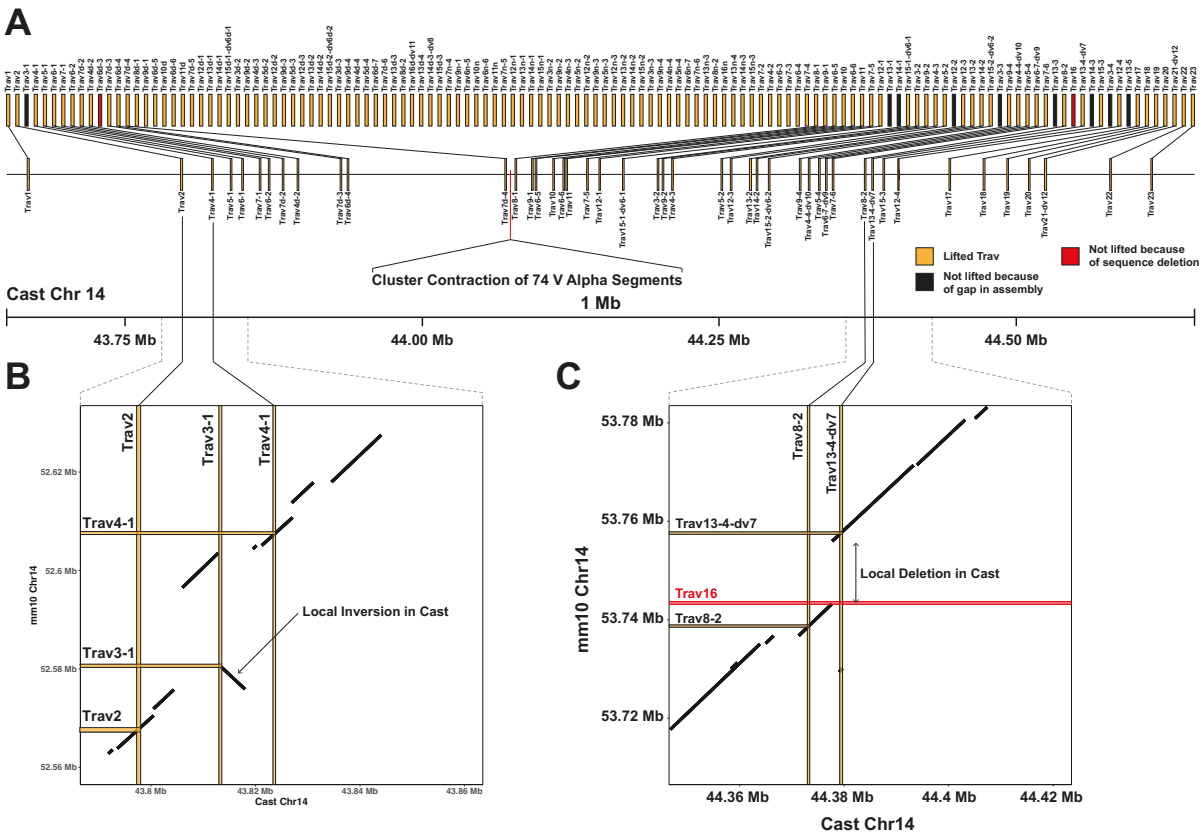


Figure 3: Comparison of the mm10 and CAST Trav gene segment loci. (A) Connected segments indicate full-length Trav genes that were lifted to the CAST genome and confirmed in a six-frame translation BLAT (45 in total). Trav gene segments that were not lifted because of gaps in the assembly (black) or are deleted in the CAST genomic sequence (red) lack a connecting line to their mm10 ortholog. **(B)** Zoom-in on the sequence surrounding *Trav3-1* indicates a local sequence inversion in the CAST assembly. **(C)** Zoom-in on the sequence surrounding *Trav16* indicates a local sequence deletion in CAST.

Gene segment usage validation by gene expression analysis

To validate our gene segment annotation and their usage in the TCR repertoire, we extracted CD8$^+$ T cells from the spleen of a 10-week-old male CAST mouse (**see methods**). We then generated TCR repertoire sequencing libraries using the Chromium Next GEM Single Cell 5' Kit. Critically, this kit utilizes a template-switch based library generation approach, such that it can recover the complete repertoire of expressed TCRs, regardless of the actual recombined 5' gene segment. Next, we assembled full-length

TCR sequences derived from sequencing reads that shared an identical cell barcode and were able to recover a total of 4535 unique Trav and 5389 unique Trbv transcripts (**see methods**). To identify V gene segment alleles, we then collapsed transcripts with identical framework region sequences. To distinguish sequencing and PCR errors from rare alleles, we required each candidate allelic variant to be observed with at least two unique CDR3 sequences. The resulting set of V gene alleles was intersected with the 45 Trav and 34 Trbv sequences generated by direct liftover of mm10 V gene segment coordinates to the CAST genome assembly. For the Trav cluster we identified all 9 full-length coding sequences for the V gene segments that were not identified from the liftover approach, presumably because of assembly gaps in the CAST genome (see previous section). Accordingly, the final set of sequences consisted of 54 Trav gene segments. For all other gene segment loci (Jα, Vβ, Dβ, Jβ) no additional sequences were identified in the transcriptomic data. Next, we assembled a full TCRα and TCRβ V(D)J reference library using the *buildLibrary* function of the MiXCR software toolkit [45]. We then used this custom species-specific library to map the sequencing reads of the CAST TCR library, resulting in 86.36% of successfully aligned reads with a V-J spanning clonotype. Next, we analyzed the V(D)J gene segment usage frequencies across all T cells to validate the expression of the identified set of gene segments (**Fig. 4A** and **Fig. 4B**). For Trav genes, we validated the expression of 42 of the 54 V gene segments included in the CAST specific V(D)J reference. The 12 Trav genes that were not expressed consisted of the 9 Trav genes which are annotated as pseudogenes in the mm10-based IMGT reference as well as *Trav13-4-dv7*, *Trav13-5* and *Trav18*, for which we have previously validated the absence of expression in C57BL/6 mice (see Chapter 3, also *Peters et al., 2024, unpublished*). In TCRα chains we observed a prominent pattern of preferential pairing of distal Vα segments with proximal Jα segments and vice versa. This pattern has been described before [46] and provides further evidence for the correct annotation of the underlying gene segments in the TCRα locus. For Traj genes we validated the expression of 44 of the 60 Traj genes included in the respective reference. All of the unexpressed Traj genes are annotated as pseudogenes or ORFs in the mm10-based IMGT reference, for which we have validated the absence of their expression in C57BL/6 (see Chapter 3,

also *Peters et al., 2024, unpublished*). In summary, we were able to confirm the expression of 42 Trav genes and 44 Traj genes in CAST mice.

In the TCRβ chains we observed the expression of all 22 Trbv genes that are annotated as functional Trbv genes in the IMGT reference as well as expression of *Trbv21* which is annotated as ORF. We can also confirm the absence of *Trbv24* expression in C57BL/6 (see Chapter 3, also *Peters et al., 2024, unpublished*) and showed that this was caused by a SNP that introduces a premature stop codon (p.Y109X) at the 3' end of the FR3 region. Critically, in the CAST *Trbv24* sequence this amino acid change was not observed, leading to the frequent utilization of this gene in expressed TCRβ chains. We also observed the expression of all Trbj gene segments that are annotated as functional genes in the IMGT reference as well as expression of *Trbj1-6* which is annotated as ORF. In summary, we were able to confirm the expression of 23 Trbv genes and 12 Trbj genes in CAST mice.

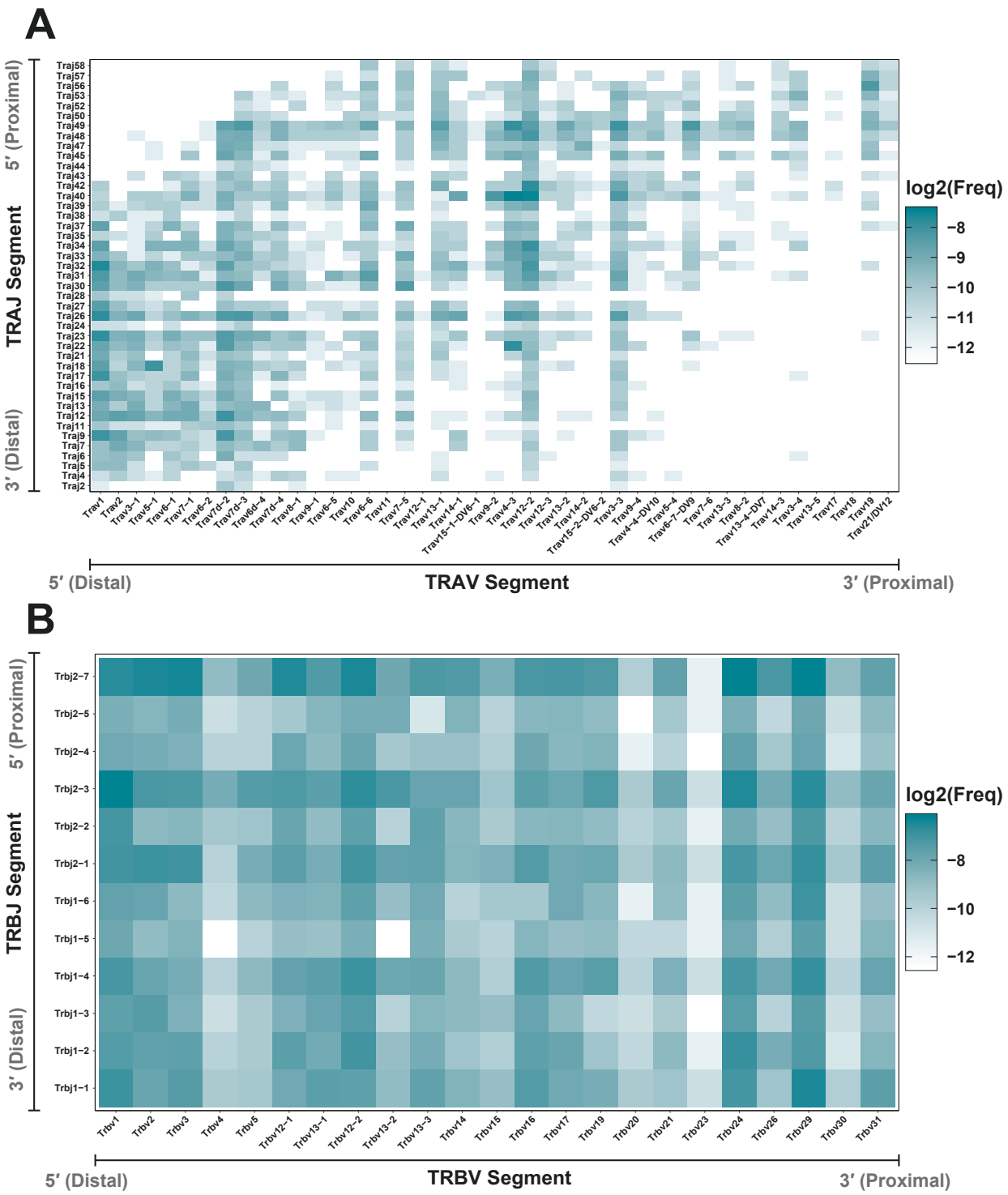


Figure 4: V-J gene segment usage frequency (log2) in the CAST CD8⁺ T cell TCR repertoire. Individual V(D)J gene segments are represented in their chromosomal order for the TCRα (A) and TCRβ (B) chain. In the TCRα chain distal V segments are more frequently paired with proximal J segments and vice versa.

Discussion

Diversification of immune receptors by somatic recombination is a key feature of the adaptive immune system. The number of available gene segments that are rearranged during V(D)J recombination to generate functional receptors varies significantly in the TCRα and TCRβ chains of different species. These differences are caused by a high frequency of rearrangements in the germline sequence of the underlying gene segment loci, which can lead to heritable locus expansions and contractions. Duplicated gene segments share extensive sequence homology and are therefore grouped into gene segment families. The murine Trav cluster has undergone a recent expansion that resulted in two duplicated blocks (the “d” and “n” blocks) which contain about two-thirds of the ancestral Trav gene segments. In this study we provide evidence for an even more recent rearrangement of the Trav cluster that has led to a major locus contraction in *Mus musculus castaneus* including the loss of 74 Trav gene segments relative to the other sub-species of *Mus musculus* (e.g. *Mus musculus musculus* and *Mus musculus domesticus*). Based on their latest common ancestor, this locus contraction is likely to have occurred less than 500,000 years ago. The frequent sequence duplications leading to highly homologous gene segment family members severely complicates the high-quality assembly of the Trav cluster in reference genomes. At those genomic regions we observed large gaps in the most recent genome assemblies of the four analyzed inbred mouse strains and showed that those gaps often overlap with the predicted location of Trav gene segments. Consequently, V gene segment inference from genomic sequences is prone to yield incomplete gene segment repertoires due to the lack of available sequence information. By utilizing transcriptomic data of CAST TCR receptors, we were able to confirm the expression of 9 Trav gene segments that we were unable to infer from the respective genomic sequences. Critically, we also identified a functional Trav gene segment (*Trav3-1*) with inverted sequence orientation. To the best of our knowledge, functional inverted V gene segments have not been reported for the TCRα chain and have previously only been observed in the form of the highly conserved inverted Trbv gene segment (e.g. murine *Trbv31* and human *TRBV30*) at the 3’ end of the TCRβ locus. Based on pairwise sequence alignments, we showed that large sequence inversions are also present in the Trav clusters of other inbred mouse strains (e.g. SPRET) and therefore

likely depict a common feature of rearranged TCR loci that can contain functional gene segments.

In contrast, the remaining gene segment clusters (Jα, Vβ, Dβ, Jβ) showed significantly less major sequence rearrangements across the four different inbred mouse strains. Inference of the CDS of those gene segments from genomic sequences resulted in highly similar, and in most cases identical numbers of predicted functional gene segments across all four mouse species/strains. In line with these predictions, we were able to confirm the expression of all Jα, Vβ, Dβ and Jβ gene segments that are annotated as functional in the mm10-based IMGT V(D)J reference database.

Based on our results, we hypothesized the Trav cluster, relative to other gene segment clusters, is evolutionarily favored to undergo frequent rearrangements, leading to cluster expansions and contractions. An excess of Trav gene segments relative to Trbv gene segments is observed in the majority of mammalian species alongside large numbers of Traj gene segments [18]. The temporally ordered generation of TCRs is initiated by TCRβ rearrangements, a process that is stringently controlled by restricted *Rag* expression and allelic exclusion. A specific checkpoint, termed the β-checkpoint, ensures that only T cells with a functional TCRβ chain progress to the DP maturation stage. In contrast, at the TCRα locus rearrangement is far less stringent with limited allelic exclusion and prolonged *Rag* expression leading to continuous rearrangements of TCRα chains over an extended period of time. The ability to “test” different TCRα rearrangements in combination with a pre-defined TCRβ during thymic selection, should evolutionarily favor extended periods of *Rag* expression and a larger set of Trav and Traj gene segments. Additionally, we can show that thymic selection is more likely to reject particular Trbv compared to Trav gene segments, based on their affinity to MHCs of different MHC-haplotypes (see Chapter 3, also *Peters et al., 2024, unpublished*).

It is therefore likely that purifying selection is less strong for TCRα gene segments, and in fact it may be that expansion of V segments may allow more T cells to survive thymic selection, thus contributing to adaptive immunity. Under such a scenario, even severe rearrangements in the germline configuration of these loci can persist. Following this line of arguments, the initially assembled TCRβ chains could be under selective pressure to maintain a baseline TCR functionality (e.g. by showing appropriate MHC affinity), while

the TCRα chains exhibit greater flexibility which can facilitate the rapid adaptation to the exposure of varying pathogens.

In this study, we have highlighted the immense diversity of Trav gene segments that can be observed even in closely related species. We showed that utilizing available genomic sequences of model organisms to predict the sequence of these gene segments often yields incomplete repertoires. This is mainly caused by the dynamic changes in the underlying loci including duplications, contractions and inversions which collectively result in frequent assembly gaps for these regions. Because V(D)J gene segments are the building blocks of functional TCRs, variance in available segments should have significant impact on the TCR diversity of an individual. Unraveling the fine-scale structure of TCR loci is therefore crucial to investigate the evolution and functional specifics of adaptive immune systems.

Materials & Methods

Mice

All mice were housed in the animal facility of the Friedrich-Miescher Laboratory of the Max-Planck Society. Experiments were performed under license issued by the local competent authority (EB 01/21 M). Mice were originally bought from Charles River Laboratories (Sulzfeld, Germany). Spleens were collected from mice aged 9-11 weeks. The following mouse strains were used in the experiments: C57BL/6J (The Jackson Laboratory, Strain #: 000664), CAST/EiJ (The Jackson Laboratory, Strain #: 000928).

Isolation of CD8a$^+$ T-cells

Spleens of euthanized mice were collected and placed on a 40µm cell-strainer. Spleens were then pressed through the strainer using the backside of a syringe plunger. After thorough rising of the cell-strainer using ice-cold PBS, the flow-through was centrifuged at 400xg 4°C for 10 minutes in a swing-bucket centrifuge. Afterwards, supernatant was carefully discarded, and the cell pellet was resuspended in 1ml ice-cold PBS + 2% FBS. Isolation of CD8a$^+$ T-cells was then done using the “Dynabeads™ FlowComp™ Mouse CD8 Kit” (Invitrogen, 11462D) according to the manufacturer’s instructions.

Single-cell TCR sequencing library preparation

After isolation of CAST T cells, TCR sequencing libraries were generated using the 10x Genomics Immune Profiling platform (Chromium Next GEM Single Cell 5' Kit v2) according to the manufacturer's instructions. T cells were processed in two separate reactions (two wells of a 10x chip), each with 2.500 input cells. V(D)J sequencing libraries were sequenced at 5.000 reads/cell. Sequencing was done on the Nova-seq 6000 platform by Illumina using S4 2x150bp v1.5 kits with the following sequencing-cycle set-up: R1: 150 cycles, i7 index: 10 cycles, i5 index: 10 cycles, R2: 150 cycles.

TCR sequencing data processing

Raw fastq-files were processed using the *cellranger vd*j software toolkit provided by 10x Genomics with the built-in mm10 based VDJ-reference (GRCm38-ensemble-7.0.0). In this pipeline fragmented reads are combined into full length contigs based on sequence overlaps in reads and matching cellular barcodes. Importantly, high-quality base call polymorphisms relative to the provided V(D)J reference remain unmodified, so that the generated *filtered_contig.fastq* files contain species-specific allelic variants of these gene segments.

Species-specific V(D)J reference libraries

The generated *filtered_contig.fastq* files were directly passed to the MiXCR alignment step ("*align*", --species mmu, --preset generic-amplicon --floating-left-alignment-boundary --floating-right-alignment-boundary C --rna) to generate binary *vdjca*-files. We then used *mixcr exportAlignments* (--dont-impute-germline-on-export -allNFeatures UTR5Begin FR3End) to extract gene-features so that SNPs in candidate-alleles are not modified to match the provided reference. For each candidate V(D)J-allele we then used the extremely unique combination of associated UMI and CDR3 sequences to distinguish low-frequency alleles from alleles generated by sequencing or PCR errors by requiring each allele to be identified with at least two unique CDR3/UMI combinations. The list of identified V, D and J segment alleles was then used to generate a MiXCR compatible reference libraries for each species using the *buildLibrary* function implemented in MiXCR. Since the underlying RNA-based input libraries are generated using template-

switching rather than multiplex-PCR, they allow for the discovery of *de novo* V(D)J-segments since template-switch based cDNA libraries do not require previous knowledge of the entire set of gene-segments for amplification.

Alignment of sequencing reads using MiXCR

Raw fastq-files containing TCR sequencing reads were integrated into a custom MiXCR pipeline (MiXCR version 4.5.0) using the following steps:

1) *mixcr align*
```
-- preset generic-ht-single-cell-amplicon-with-umi
-- library Species Specific custom library (see above)
-- tag-pattern ^(CELL:N(16))(UMI:N(10))(R1:*)\^(R2:*)
-- floating-left-alignment-boundary
-- floating-right-alignment-boundary C
-  OvParameters.geneFeatureToAlign=VRegionWithP
-  OminSumScore=100
```

2) *mixcr refineTagsAndSort*

3) *mixcr assemble*
```
-- assemble-clonotypes-by CDR3
-- cell-level
```

We then used *mixcr exportClones* to extract the required information for all downstream analysis (e.g., cellular barcodes, transcript counts, V(D)J segments, CDR3 amino acid and nucleotide sequence etc.).

Reference genome assemblies

All assembled murine reference genomes were received from the Ensemble database (release 102). The following reference genomes were used: mus_musculus_129s1svimj, mus_musculus_pwkphj, mus_musculus_casteij, mus_spretus and the standard GRCm38 (mm10) mouse reference genome.

Pairwise alignment of genomic sequences

We performed a local pairwise alignment of genomic sequences of the TCR loci across all analyzed mice by using minimap2 [47] with the following parameters: “-PD -k19

-w19 -m200 -t48". The resulting pairwise alignment files (.paf) were then used to plot alignment dotplots using the R package *pafr* [48].

Liftover of gene coordinates and genome track visualization

Coordinates of the annotated V(D)J gene segments in the GRCm38/mm10 genome were lifted to the genome assembly of the alternative mouse strains using GTF files downloaded from the Ensemble database (e.g. Mus_musculus.GRCm38.102.gtf) and the corresponding "UCSC Chain Files" (e.g. mm10ToGCA_001624445.1.over.chain.gz). The generated GTF files contained the chromosomal locations of the lifted gene segments. These locations were used to generate bed-interval files that were visualized using the Integrative Genomics Viewer [49].

Acknowledgements

We thank all past and present members of the Chan and Jones laboratory for input into experimental design, helpful discussion and improving the manuscript. We especially thank Felicity Jones for her scientific input throughout the entire study. We thank Sinja Mattes and the remaining team of animal caretakers at the Friedrich-Miescher Laboratory led by Cemal Yilmaz. M.P. and D.S. are supported by an International Max Planck Research School fellowship. Y.F.C is supported by the European Research Council Starting Grant 639096 "HybridMix" and Proof-of-Concept Grant 101069216 "Haplotagging". The research done in this study is supported by the Max Planck Society.

Author Contributions

M.P. and Y.F.C. designed the experiments. M.P. performed all experiments with support of V.S.. M.P. and Y.F.C performed the computational analysis. M.P. and Y.F.C wrote the manuscript. V.S., D.S. and Y.F.C. provided support for the experiments and the computational analysis. All authors reviewed the manuscript. Y.F.C. direct the study.

Declaration of Interest

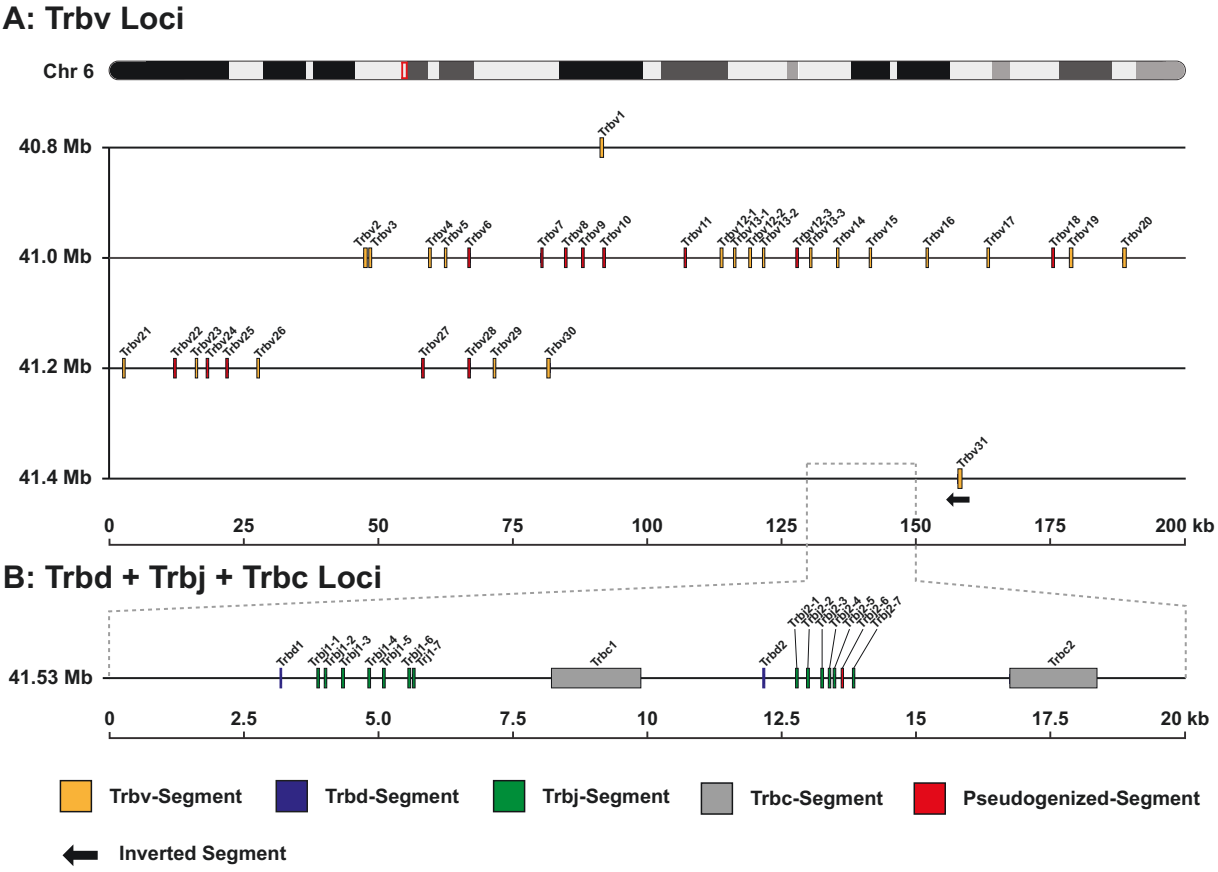
The authors declare no competing interests.

Literature

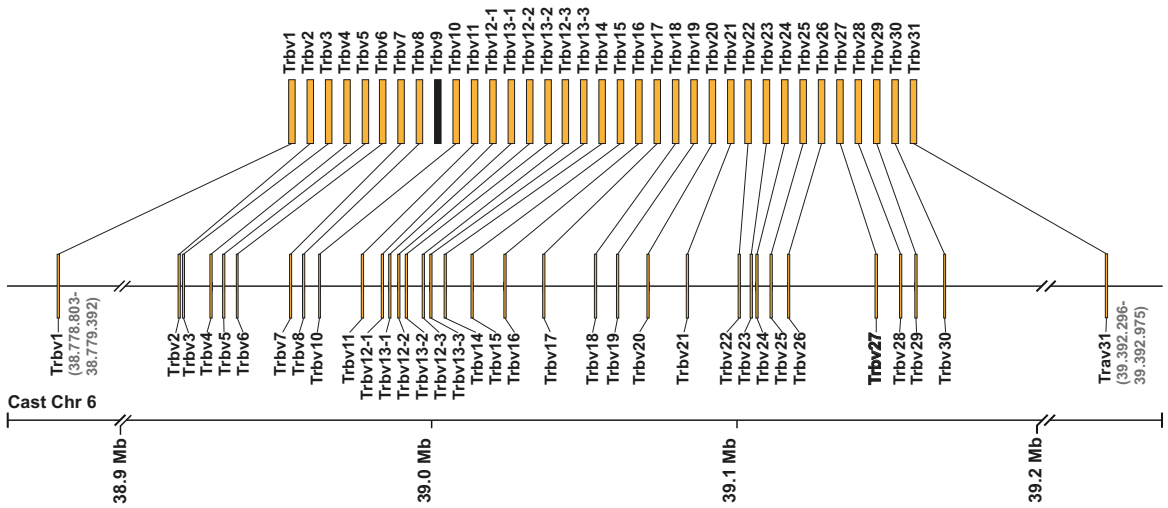
1. Hozumi, N. and S. Tonegawa, *Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions.* Proc Natl Acad Sci U S A, 1976. **73**(10): p. 3628-32.
2. Davis, M.M. and P.J. Bjorkman, *T-cell antigen receptor genes and T-cell recognition.* Nature, 1988. **334**(6181): p. 395-402.
3. Thierry Mora, A.W., *Quantifying lymphocyte receptor diversity.* Systems Immunology, (ed. J. Das, C. Jayaprakash), 2019.
4. Doherty, P.C., J.M. Riberdy, and G.T. Belz, *Quantitative analysis of the CD8 T-cell response to readily eliminated and persistent viruses.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2000. **355**(1400): p. 1093-1101.
5. Arstila, T.P., et al., *A direct estimate of the human αβ T cell receptor diversity.* Science, 1999. **286**(5441): p. 958-961.
6. Oettinger, M.A., et al., *Rag-1 and Rag-2, Adjacent Genes That Synergistically Activate V(D)J Recombination.* Science, 1990. **248**(4962): p. 1517-1523.
7. Taccioli, G.E., et al., *Impairment of V(D)J recombination in double-strand break repair mutants.* Science, 1993. **260**(5105): p. 207-10.
8. Komori, T., et al., *Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes.* Science, 1993. **261**(5125): p. 1171-5.
9. Max, E.E., J.G. Seidman, and P. Leder, *Sequences of five potential recombination sites encoded close to an immunoglobulin kappa constant region gene.* Proc Natl Acad Sci U S A, 1979. **76**(7): p. 3450-4.
10. van Gent, D.C., D.A. Ramsden, and M. Gellert, *The RAG1 and RAG2 proteins establish the 12/23 rule in V(D)J recombination.* Cell, 1996. **85**(1): p. 107-13.
11. Du Pasquier, L., I. Zucchetti, and R. De Santis, *Immunoglobulin superfamily receptors in protochordates: before RAG time.* Immunological Reviews, 2004. **198**: p. 233-248.
12. Lefranc, M.P., et al., *IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains.* Developmental and Comparative Immunology, 2003. **27**(1): p. 55-77.
13. Chlewicki, L.K., et al., *High-affinity, peptide-specific T cell receptors can be generated by mutations in CDR1, CDR2 or CDR3.* Journal of Molecular Biology, 2005. **346**(1): p. 223-239.
14. Engel, I. and S.M. Hedrick, *Site-Directed Mutations in the Vdj Junctional Region of a T-Cell Receptor Beta-Chain Cause Changes in Antigenic Peptide Recognition.* Cell, 1988. **54**(4): p. 473-484.
15. Jorgensen, J.L., et al., *Mapping T-Cell Receptor Peptide Contacts by Variant Peptide Immunization of Single-Chain Transgenics.* Nature, 1992. **355**(6357): p. 224-230.
16. Olivieri, D., et al., *An automated algorithm for extracting functional immunologic V-genes from genomes in jawed vertebrates.* Immunogenetics, 2013. **65**(9): p. 691-702.
17. Olivieri, D.N., et al., *Genomic V exons from whole genome shotgun data in reptiles.* Immunogenetics, 2014. **66**(7-8): p. 479-92.
18. Olivieri, D.N., S. Gambon-Cerda, and F. Gambon-Deza, *Evolution of V genes from the TRV loci of mammals.* Immunogenetics, 2015. **67**(7): p. 371-84.
19. Parra, Z.E., et al., *Comparative genomic analysis and evolution of the T cell receptor loci in the opossum Monodelphis domestica.* BMC Genomics, 2008. **9**: p. 111.
20. Parra, Z.E., et al., *A unique T cell receptor discovered in marsupials.* Proc Natl Acad Sci U S A, 2007. **104**(23): p. 9776-81.
21. Glusman, G., et al., *Comparative genomics of the human and mouse T cell receptor loci.* Immunity, 2001. **15**(3): p. 337-349.
22. Charlemagne, J., et al., *T-cell receptors in ectothermic vertebrates.* Immunol Rev, 1998. **166**: p. 87-102.
23. Sitnikova, T. and C. Su, *Coevolution of immunoglobulin heavy- and light-chain variable-region gene families.* Molecular Biology and Evolution, 1998. **15**(6): p. 617-625.
24. Migalska, M., A. Sebastian, and J. Radwan, *Profiling of the TCRβ repertoire in non-model species using high-throughput sequencing.* Scientific Reports, 2018. **8**.

25. Nei, M., X. Gu, and T. Sitnikova, *Evolution by the birth-and-death process in multigene families of the vertebrate immune system.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(15): p. 7799-7806.
26. Nei, M., *Gene Duplication and Nucleotide Substitution in Evolution.* Nature, 1969. **221**(5175): p. 40-+.
27. Gojobori, T. and M. Nei, *Concerted Evolution of the Immunoglobulin-Vh Gene Family.* Molecular Biology and Evolution, 1984. **1**(2): p. 195-212.
28. Hughes, A.L. and M. Nei, *Evolution of the Major Histocompatibility Complex - Independent Origin of Nonclassical Class-I Genes in Different Groups of Mammals.* Molecular Biology and Evolution, 1989. **6**(6): p. 559-579.
29. Ota, T. and M. Nei, *Divergent Evolution and Evolution by the Birth-and-Death Process in the Immunoglobulin V-H Gene Family.* Molecular Biology and Evolution, 1994. **11**(3): p. 469-482.
30. Clark, S.P., et al., *Comparison of human and mouse T-cell receptor variable gene segment subfamilies.* Immunogenetics, 1995. **42**(6): p. 531-40.
31. Su, C. and M. Nei, *Evolutionary dynamics of the T-cell receptor VB gene family as inferred from the human and mouse genomic sequences.* Molecular Biology and Evolution, 2001. **18**(4): p. 503-513.
32. Brown, D.D., P.C. Wensink, and E. Jordan, *A comparison of the ribosomal DNA's of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes.* J Mol Biol, 1972. **63**(1): p. 57-73.
33. Keane, T.M., et al., *Mouse genomic variation and its effect on phenotypes and gene regulation.* Nature, 2011. **477**(7364): p. 289-94.
34. She, J.X., et al., *Molecular Phylogenies in the Genus Mus - Comparative-Analysis of Electrophoretic, Scndna Hybridization, and Mtdna Rflp Data.* Biological Journal of the Linnean Society, 1990. **41**(1-3): p. 83-103.
35. Suzuki, H., et al., *Temporal, spatial, and ecological modes of evolution of Eurasian based on mitochondrial and nuclear gene sequences.* Molecular Phylogenetics and Evolution, 2004. **33**(3): p. 626-646.
36. Guénet, J.L. and F. Bonhomme, *Wild mice: an ever-increasing contribution to a popular mammalian model.* Trends in Genetics, 2003. **19**(1): p. 24-31.
37. Lilue, J.T., et al., *Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci.* Nature Genetics, 2018. **50**(11): p. 1574-+.
38. Abolins, S., et al., *The comparative immunology of wild and laboratory mice.* Nature Communications, 2017. **8**.
39. Abolins, S.R., et al., *Measures of immune function of wild mice.* Molecular Ecology, 2011. **20**(5): p. 881-892.
40. Lefranc, M.P., et al., *IMGT, the international ImMunoGeneTics information system.* Nucleic Acids Res, 2009. **37**(Database issue): p. D1006-12.
41. Zhang, T., et al., *Genomic organization of the chicken TCRbeta locus originated by duplication of a Vbeta segment combined with a trypsinogen gene.* Vet Immunol Immunopathol, 2020. **219**: p. 109974.
42. Olivieri, D.N. and F. Gambón-Deza, *Iterative Variable Gene Discovery from Whole Genome Sequencing with a Bootstrapped Multiresolution Algorithm.* Computational and Mathematical Methods in Medicine, 2019. **2019**.
43. Sirupurapu, V., Y. Safonova, and P.A. Pevzner, *Gene prediction in the immunoglobulin loci.* Genome Res, 2022. **32**(6): p. 1152-1169.
44. Goios, A., et al., *mtDNA phylogeny and evolution of laboratory mouse strains.* Genome Res, 2007. **17**(3): p. 293-8.
45. Bolotin, D.A., et al., *MiXCR: software for comprehensive adaptive immunity profiling.* Nat Methods, 2015. **12**(5): p. 380-1.
46. Kitaura, K., et al., *A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) alpha and beta repertoires and identifying potential new invariant TCR alpha chains.* BMC Immunol, 2016. **17**(1): p. 38.
47. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.
48. Winter, D., *Read, Manipulate and Visualize 'Pairwise mApping Format' Data.* Package ‘pafr’, 2020.
49. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

Supplement



Supplementary Figure 1: Genomic locations of V/D/J gene segments in the TCRβ loci as annotated in the GRC38/mm10 based IMGT annotation. (A) TCRβ variable genes (35 total Trav segments) are located in an 800 kb window on chromosome 6. *Trbv31* is located upstream of the D/J/C clusters in an inverted sequence orientation. **(B)** The D/J/C loci consist of two blocks of a single D gene segment 7 J gene segments and a constant region located upstream of *Trbv1-Trbv30* in a 20 kb window.



Supplementary Figure 2: Comparison of the mm10 and CAST Trbv gene segment loci. Trbv gene segments that were lifted to the CAST genome at full-length are connected to their mm10 ortholog. Similar to their location in mm10 *Trbv1* is located downstream and *Trbv31* is located upstream of main Trbv cluster. A homologous sequence of the mm10 *Trbv9* pseudogene could not be identified on chromosome 6 of CAST.

Chapter 2: Genetic determinants of distinct CD8⁺ α/β-TCR repertoires in the genus *Mus*

Moritz Peters[1*], Volker Soltys[1], Dingwen Su[1], Marek Kučka[1,2], Yingguang Frank Chan[1,3*]

1. Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany
2. Department of Translational Genomics, University of Cologne, 50931 Cologne, Germany
3. University of Groningen, Groningen Institute for Evolutionary Life Sciences, 9747 AG Groningen, Netherlands

*Corresponding authors: moritz.peters@tuebingen.mpg.de, frank.chan@rug.nl

Status in submission process: Advanced manuscript; ready for submission

Abstract

The adaptive immune system's efficacy relies on the diversity of T cell receptors and the ability to distinguish between self and foreign antigens. Analysis of the paired heterodimeric αβ-TCR chains of individual T cells requires single-cell resolution, but existing single-cell approaches offer limited coverage of the vast TCR repertoire diversity. Here we introduce CITR-seq, a novel, instrument-free, high-throughput method for single-cell TCR sequencing with >88% αβ-TCR pairing precision.

We analyzed the TCR repertoires of CD8⁺ T cells originated from 32 inbred mice using CITR-seq, comprising four evolutionary divergent sister species and their F1 hybrids. Overall, we identified more than 5 million confidently paired TCRs. We found that V(D)J gene usage patterns are highly specific to the genotype and that Vβ-gene usage is strongly impacted by thymic selection. Using F1 hybrids, we show that differences in gene segment usage across species are likely caused by *cis*-acting factors prior to thymic selection, which imposed strong allelic biases. At the greatest divergence, this led to increased rates of TCR depletion through rejection of particular Vβ-genes. TCR repertoire overlap analysis across all mice revealed that sharing of identical paired CDR3 amino acid motifs is four times more frequent than predicted by random pairing of TCRα and TCRβ chains, with significantly increased sharing rates among related individuals. Collectively, we show that beyond the stochastic nature of TCR repertoire generation, genetic factors contribute significantly to the shape of an individual's repertoire.

Introduction

Adaptive immunity relies on the recognition of antigens presented on the surface of virtually all nucleated cells through class I and II major histocompatibility complexes (MHC). These MHC-bound antigens are recognized by T cell receptors (TCRs) expressed on the surface of T cells, which collectively possess the remarkable ability to discriminate between antigens of “self” and “foreign” origin. This is critical for preserving self-tolerance, thereby preventing autoimmunity, while also enabling the identification of pathogen-infected or malignant cells to initiate an immune response [1]. The nature of an immune response depends largely on whether an antigen is presented via class I or class II MHC complexes, which are targeted by $CD8^+$ cytotoxic T cells [2] or $CD4^+$ helper T cells [3], respectively. The former can induce apoptosis in targeted cells while the latter can trigger secondary immune cascades involving B-lymphocytes and cells of the innate immune system. In both cases, TCRs are the key molecules that mediate signaling and enable a broad spectrum of immune responses.

TCRs are primarily composed of two heterodimeric chains, TCRα and TCRβ, both of which arise from somatic rearrangements of gene segments during T cell development. This rearrangement process, known as V(D)J recombination, generates diversity through joining variable (V), diversity (D, exclusive to TCRβ) and joining (J) gene segments to a constant region, thereby generating a unique TCR receptor in each individual T cell [4]. Additional diversity is introduced through nucleotide insertions and deletions at each junction during V(D)J recombination [5]. In the expressed TCRα and TCRβ chains, the resulting highly polymorphic junctional region is situated in closest proximity to antigens presented by MHCs to serve as a binding pocket [6] and is termed complementarity-determining region 3 (CDR3). The other CDRs, 1 and 2, constitute germline encoded regions within the V-segments of TCR chains and are believed to primarily facilitate TCR-MHC binding and are less relevant for antigen recognition [7, 8] (**Fig. 1A**).

The antigen specificity of each rearranged TCR as well as its affinity to MHCs is evaluated in a key multi-step process called thymic selection. It takes place during T cell maturation which is generally classified by the intra-thymic progression from the CD8/CD4 double negative (DN) stages of lymphoid precursors to the single positive stage (SP) of mature

T cells. The selection process is initiated at the so called β-checkpoint [9], at which the successful rearrangement of an in-frame TCRβ chain is controlled. Afterwards, the fully assembled αβ-TCR is tested for its affinity to MHCs during positive selection and its specificity to presented self-antigens during negative selection. Both processes are chronologically and spatially separated: Positive selection occurs in the cortex through interactions with cortical thymic epithelial cells (cTECs), whereas the subsequent negative selection occurs in the medulla through interactions with medullary thymic epithelial cells (mTECs) (reviewed here [10]). Overall, only about 5% of T cell precursors survive thymic selection of their TCRs by demonstrating adequate affinity to MHC complexes while simultaneously exhibiting tolerance towards the broad spectrum of presented self-antigens and thus thymic selection significantly decreases the diversity in the TCR repertoire (**Fig. 1B**).

In our current understanding, nucleotide insertions and deletions, V(D)J-segment usage and αβ-TCR pairing are mostly seen as stochastic events that give rise to highly unique and dynamic TCR repertoires within and across individuals [11]. However, recent work increasingly suggests that the diversity of TCR repertoires also relies on genetically encoded differences across individuals [12-15]. For example, V-segment usage in identical twins exhibits much greater similarity compared to unrelated individuals [16]. Notably, this provides evidence that genetics may operate at two different levels: V(D)J recombination as well as thymic selection, as indicated by the fact that identical twins also share the same set of MHC class I and II (also known as human leukocyte antigen, or HLA) alleles. This observation, and the multi-level genetic determinants that collectively shape TCR diversity, are the focus of at least two debates: one concerning the existence of a co-evolutionary feedback process between TCR and MHC binding [17-21], and the other on whether MHC heterozygosity is evolutionarily optimal due to the presentation of a broader immunopeptidome [22, 23], or alternatively, deleterious due to a high frequency of presented self-peptides, leading to increased depletion of autoreactive TCRs [24, 25]. The extreme diversity of both binding partners, TCR and MHC, makes answering these questions extremely challenging. By contrast, panels of inbred mouse lines spanning within- and across-species diversity, along with their F1 hybrids, provide a tractable setup

to address the question regarding the role of MHC alleles in shaping repertoire diversity during thymic selection.

Estimates on the theoretical αβ-TCR diversity vary greatly by species as well as methodology. Initial theoretical estimates on αβ-TCR diversity were approximately 10^{15} in mice [4] and 10^{18} [26] – 10^{20} [27] in humans. More recent calculations now greatly exceed those estimates and range up to 10^{61} [28]. However this needs context: the number of theoretical αβ-TCRs vastly outnumbers the actually realized TCRs in the repertoire of an individual, primarily because the number of present T cells of an individual (10^{12} [29] in humans and 10^{8} [30] in mice) is several orders of magnitude smaller at any given time.

Interestingly, despite the great difference in number of total T cells across different species (e.g. more than 1000x more T cells in humans than in mice), the diversity within the realized TCR repertoire has shown to be much more similar across species [30, 31]. This observation gave rise to the idea of a minimally required repertoire size defined as a functional unit of the “protecton” which is simply multiplied in species with larger numbers of total T cells [32]. Experimental validation of TCR repertoire diversity estimates still suffers from the limitations of current methodologies. While bulk assays can now feasibly analyze entire repertoires across many individuals [33-35], they leave out the critical pairing between TCRα and TCRβ chains within individual cells. Paired TCR analysis requires molecular barcoding of single cells, but existing methods often rely on pre-existing single-cell workflows, restricting analysis to thousands of T cells rather than entire repertoires [36-38]. These protocols typically utilize fluorescence-activated cell sorting (FACS), or microfluidic platforms to isolate individual cells and therefore require specialized equipment. The recent development of SPLiT-seq [39] has expanded the scope of single-cell whole transcriptome experiments to up to 10^{6} cells per experiment by using combinatorial indexing to molecularly barcode each individual cell. Despite this increase in throughput, the associated sequencing cost still substantially limits the feasibility of these methods for assessing large TCR repertoires, especially across multiple individuals.

Here, we investigate repertoires of paired αβ-TCRs from cytotoxic $CD8^{+}$ T cells by developing a targeted TCR sequencing protocol called CITR-seq (**C**ombinatorial **I**ndexing **T** cell **R**eceptor sequencing) to analyze TCR repertoires at low cost and large scale. We

apply CITR-seq to 32 individual mice from 4 distinct inbred sister species (C57BL/6J, CAST/EiJ, PWD/PhJ, SPRET/EiJ, abbreviated as BL6, CAST, PWD and SPRET, respectively) and their F1 hybrids with BL6, spanning an evolutionary divergence of approximately 3 million years [40-42] (**Fig. 1C**). The diverged but controlled genetic backgrounds provide a unique opportunity to determine the respective impact of TCR locus structure, the V, D and J gene segment usage frequency, TCR/MHC allele co-evolution via thymic selection and ultimately the joint effects on CDR3 diversity (**Fig. 1D**).

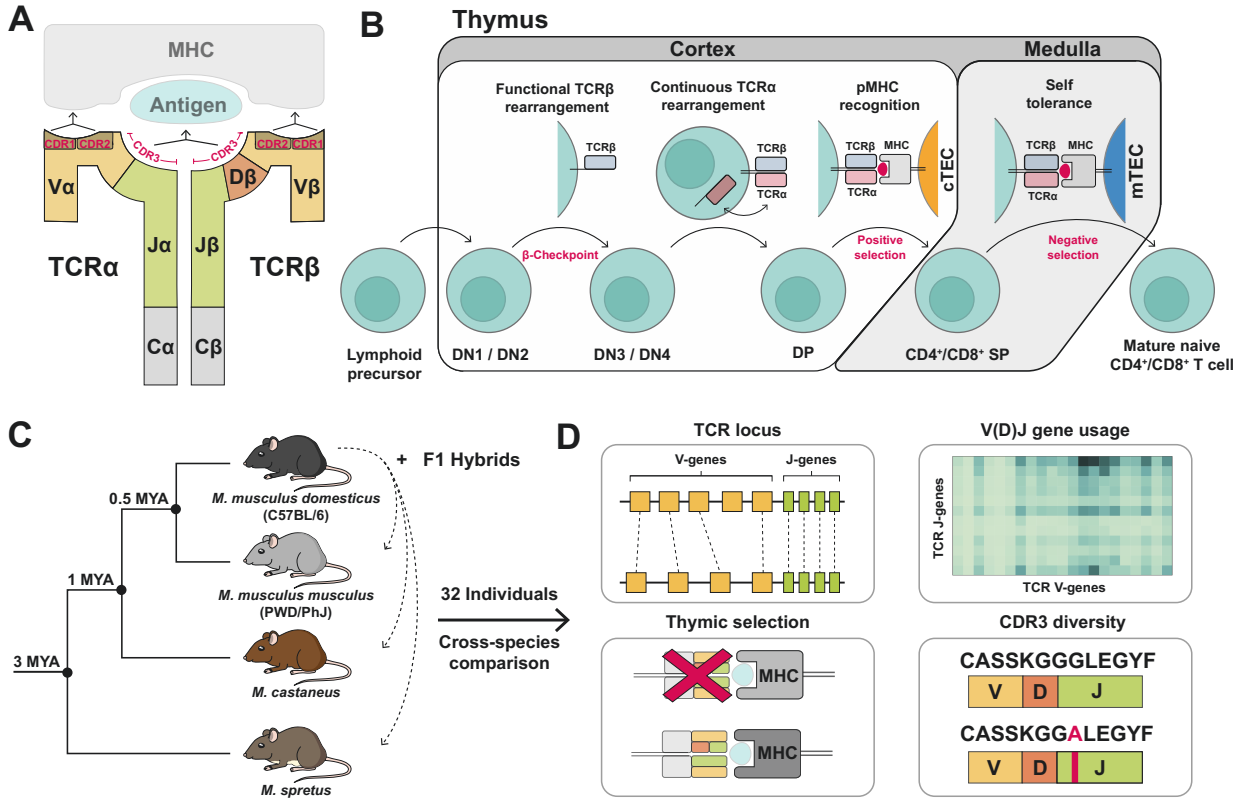


Figure 1: Introduction to T cell receptors and overview of the study design

(A) Heterodimeric αβ-TCR consisting of a V-, D- (exclusive to TCRβ) and J-gene alongside the constant (C) region. The junctional region of V(D)J genes marks the CDR3 sequence that is in the closest proximity to the antigen in the TCR-MHC complex. CDR1 and CDR2 are germline-encoded sequences of V-genes that contribute to TCR-MHC binding.

(B) T cell maturation in the thymus. Intra-thymic T cells are classified by the expression of the lineage-markers CD4 and CD8 (double negative: DN, double positive: DP and single positive: SP). T cells that successfully rearranged a functional TCRβ chain can pass the β-checkpoint. Subsequent continuous rearrangements of the TCRα chain leads to transition to the DP stage. T cells with a fully assembled αβ-TCR that is capable of binding self-MHCs on the surface of cTECs survive positive selection. Selected T cells migrate to the medulla and undergo negative selection, during which T cells that strongly bind self-MHCs on the surface of mTECs are rejected. T cells that survive both selection steps are released from the thymus.

(C) Phylogenetic tree showing the evolutionary divergence of inbred mouse species used in this study.
(D) The different aspects of TCR generation and selection analyzed in the course of this study.

Results

CITR-seq design and validation

To generate TCR repertoires we built on SPLiT-seq [39] to develop CITR-seq and modify the approach to generate RNA-based targeted paired αβ-TCR libraries (**Fig. 2A**). We first isolated $CD8^+$ T cells from spleens of 10-week-old mice by using anti-CD8 magnetic beads and subsequent purification by FACS (**Fig. S1A**). Purified $CD8^+$ T cells were either used directly or were transferred to anti-CD3 and anti-CD28 coated tissue culture plates for a 20-hour activation period before the library preparation (see **Suppl. Table 1** for detailed sample list; note that we limited the activation to 20 hours to avoid cell doubling). Primary or activated T cells were fixed and permeabilized and a set of barcoded TCRα and TCRβ constant-region primers was used to perform *in situ* reverse transcription (RT) inside individual cells (see **Suppl. Table 2** for a list of all primers and barcoding DNA-oligos). All RT-primers contain a unique molecular identifier (UMI) and a ligation overhang for single-cell barcoding. Here, cells are distributed in two split-and-pool cycles across two 96-well plates, such that the RT-primer overhangs are ligated to oligos carrying barcode segments. The split-and-pool approach allows all reactions to be performed in bulk, while giving each cell an effectively unique barcode (calculated barcode collision rate: 1.67%; **see methods**). Afterwards up to 10,000 cells are merged into sub-samples and reverse crosslinking is done to make the barcoded cDNA accessible for amplification. Second strand synthesis is done in a multiplex-PCR setup using 54 primers targeting the 5’ ends of TCR V-segments. In a final index-PCR another DNA barcode is added to each sub-sample, which expands the total barcoding space to up to 28 million unique cellular barcodes. Subsamples can then be pooled to generate the final sequencing-ready library which is compatible with standard Illumina workflows and can be further multiplexed with other sequencing libraries (**Fig. 2B**). This process is cost efficient (**see methods**) and does not require specialized instrumentation.

Using CITR-seq, we profiled TCR repertoires from a total of 9,113,392 $CD8^+$ T cells (hereafter referred to as "T cells") across all 32 individuals. Paired TCRα and TCRβ chains were successfully recovered in 75.8% of T cells, 55.4% of which carried exactly one α- and β-chain (**Fig. 2C**). To the best of our knowledge, this dataset of 5,049,334 singly α/β-paired T cells represents the largest set of paired TCRs analyzed in a single study to date (**Fig. 2D**). To assess pairing precision, we determined the rate of repeated observations of identical Vβ-Jβ-CDR3β and Vα-Jα-CDR3α mates (or "clonotypes") in a sample of 150,000 T cells that underwent clonal expansion for 72h in tissue culture. Clonal expansion through prolonged tissue culture allowed us to enrich for cells carrying the same α/β-chain pairing, the recovery of which would have been unlikely under our standard protocol. TCRβ chains that were observed at least twice in this repertoire were seen with identical TCRα chains in 88% of cells, thus representing T cells with identical clonotypes. This rate marks the lower-bound pairing precision, since with 150,000 T cells, we expect a low, but non-negligible chance of recovering the same Vβ-Jβ-CDR3β chain from two non-clonal T cells which should therefore pair with a different TCRα chain. In agreement with this high pairing precision, we observed few cells with more than two TCRα (1.4%) or TCRβ (0.5%) chains across all 32 CITR-seq samples, which is biologically implausible because of the presence of just two alleles for each chain in each T cell (**Fig. S2A**).

We then compared transcripts (UMIs) per cell counts at saturating read coverage (mean reads/cell: 184.57; **Fig. S2B**) in activated and primary T cells in CITR-seq. Activated T cells had a significantly higher UMI/cell count (14.81) compared to primary T cells (5.9, pairwise t-test; $P < 0.01$; **Fig. 2E**). We compared these values to two publicly available human TCR sequencing datasets (Parse Bioscience; **see methods**), in which 72h activated T cells also showed significantly higher average UMI per cell count (18.44 UMIs/cell) than primary T cells (4.69 UMIs/cell). As a further benchmarking effort, we generated complementary datasets for each of the four inbred mouse species from primary T cells using the Chromium Next GEM Single Cell 5' platform by 10x Genomics (**see methods**). For these, we recovered 10.1 UMIs/cell on average across samples (**Fig. 2E**).

We evaluated whether activation of T cells biases the recovered TCR repertoire by comparing V-J usage (discussed below) and clonal abundance in samples of primary and activated T cells, each down-sampled to 150,000 cells. To do so, we first compared the frequency of multiple observations of identical TCRα and TCRβ as well as full clonotypes (defined by identical V+J+CDR3 amino acid motif) across cells (**Fig. S2C**). In TCR repertoires of primary T cells and T cells that were activated for 20h, most αβ-TCR pairs were exclusive to a single cell (93% and 96.7% respectively). This is in contrast to αβ-TCR pairs in TCR repertoires of cells that were activated for 72 hours, in which less than half (48%) of pairs are exclusive to a single cell with all other pairs being observed multiple times. We therefore conclude, that in agreement with previous studies [79] the 20h activation protocol did exclude clonal expansion of T cells.

Validation of CITR-seq against 10x Genomics commercial platform

To validate complete coverage of all the functional V/J genes, we compared V-J gene usage frequencies from data generated using CITR-seq and 10x Genomics Immune Profiling. We find high correlation of V-J usage frequencies (Pearson: BL6 r = 0.91) across both methods (**Fig. S3C**). Additionally, the highly correlated V-J usage frequencies provide further evidence for the unbiased repertoire representation of 20h activated (CITR-seq) compared to primary (10x Genomics Immune Profiling) T cells.

To evaluate the coverage of cross-species repertoires, we analyzed CDR3 amino acid motif diversity in α- and β-chains both individually and jointly (**Fig. 2F**). In the 5,049,334 T cells from across all 32 samples, we detect 719,976 (14.26%) unique TCRα and 1,725,631 (34.18%) unique TCRβ chains. If analyzed jointly, 95.6% of these (4,826,991) represent unique αβ-TCR pairs. In contrast, in 9,445 paired T cells in our Chromium Next GEM Single Cell 5' datasets, we found 85% and 94.9% unique TCRα and TCRβ chains, respectively (n = 8,021 and 8,963), and nearly all (98.6%, or 9,313 T cells) represent unique αβ-TCR pairs (**Fig. S2D**). Taken together, we interpret this data to show the remarkable diversity, especially across paired TCRs: even with the throughput of CITR-seq at 5 million cells, we were not close to sampling T cell clonotypes to saturation, let alone using much more limited platforms. This further emphasizes the need for high-throughput methods to gain reasonable insight into the diversity of TCR repertoires.

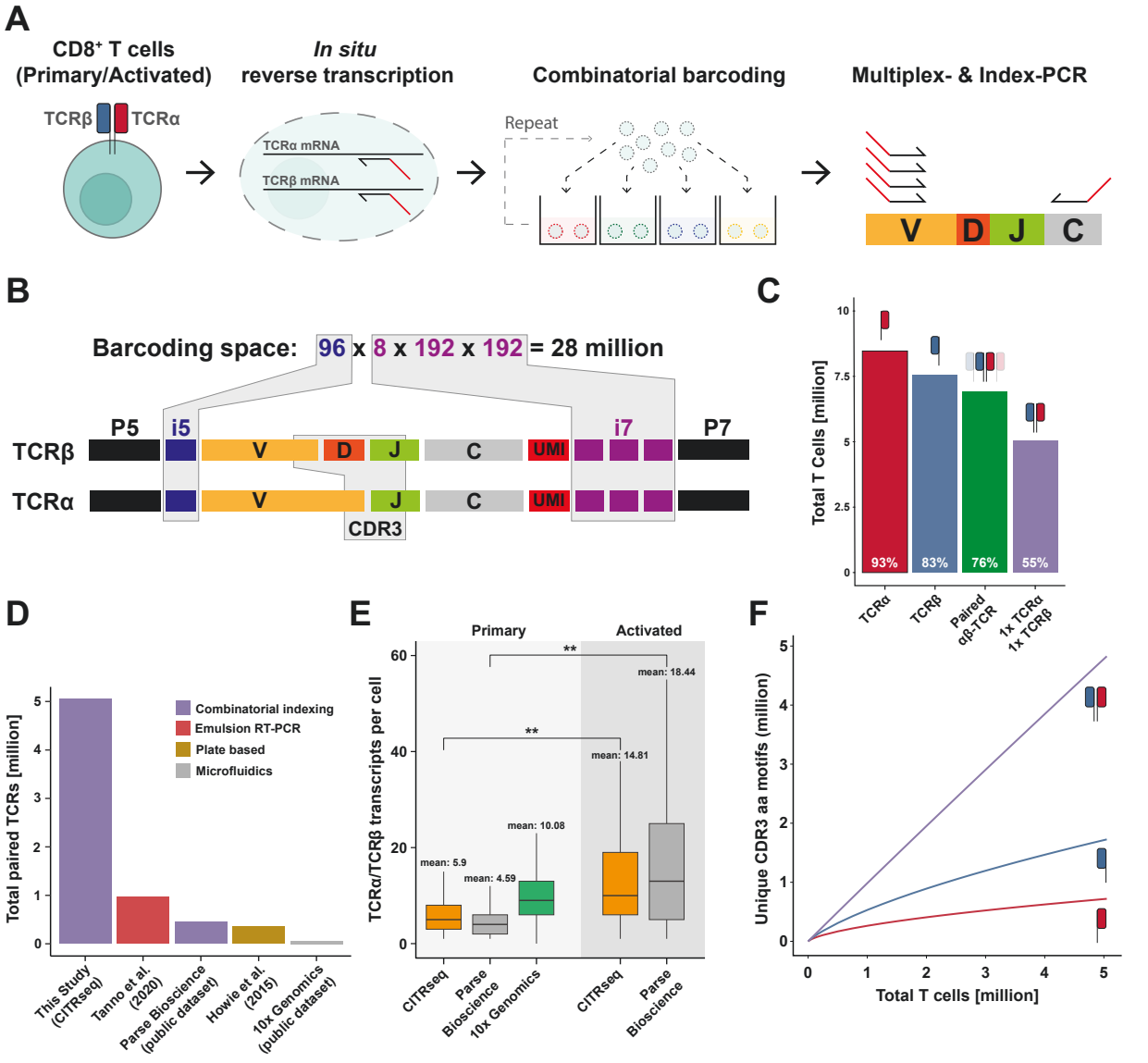


Figure 2: CITR-seq allows for the analysis of millions of confidently paired αβ-TCRs

(A) Workflow for generating paired αβ-TCR sequencing libraries using CITR-seq. Isolated T cells are fixed and permeabilized. TCRα and TCRβ are *in-situ* reverse transcribed using barcoded primers targeting both TCR constant regions. Afterwards, T cells are distributed across 96-well plates and well-specific barcodes are ligated to the cDNA. This process is repeated once by pooling all T cells and redistributing them to a second set of barcoding plates. T cells are then pooled again and split into sub-samples before reverse crosslinking. Second strand cDNA is generated in a multiplex-PCR with primers targeting the 5' region of Vα- and Vβ-genes. In a final Index-PCR a fourth barcode is added.

(B) Barcode and sequencing adapter structure in CITR-seq libraries. The different combinations of all four barcodes provide a barcoding space of more than 28 million possible barcodes. Sequencing reads fully cover CDR3α and CDR3β sequences.

(C) Pairing rate across all CITR-seq samples in this study. Fraction of the 9.113.392 total T cells that were assigned to a TCRα (93%, red), a TCRβ (83%, blue), at least one TCRα and TCRβ (76%, green) or exactly one TCRα and TCRβ chain (55%, violet).
(D) Total number of paired αβ-TCRs analyzed in different studies and publicly available datasets generated with different methods. Emulsion RT-PCT [14], PairSeq (plate-based) [43], two publicly available datasets generated with combinatorial indexing (Parse Bioscience) [44] and microfluidics (10x Genomics) [45].
(E) Mean number of TCRα and TCRβ transcripts (UMIs) per cell-barcode in primary and activated T cells in CITR-seq samples (primary and 20h activated mouse T cells), across all 10x Genomics Single-Cell Immune Profiling libraries (primary mouse T cells) and in two publicly available datasets from Parse Bioscience (primary and 72h activated human T cells) [44, 46].
(F) Total number of unique CDR3α, CDR3β or paired CDR3αβ amino acid motifs relative to the number of T cells across all 32 CITR-seq samples.

Distinct V-J usage patterns across mouse species

To compare V-J segment usage across the different mouse species, we first constructed species-specific V(D)J-segment references (**see methods**). Across all samples, mapping against the corresponding species-specific reference showed a slight increase in the total number of successfully aligned sequencing reads (PWD: +0.07%, CAST: +0.07% and SPRET: + 0.1% total reads; **Fig. S3A**) and per segment alignment scores (data not shown), relative to mapping against an mm10-based V(D)J reference provided in the MiXCR software [47]. Local alignment of TCR loci to the mm10 reference genome (GRCm38/mm10) revealed one-to-one orthology in Vβ-, Jβ- and Jα-segments. In contrast, we found extensive rearrangements, including inversions and gene-cluster triplications in the Vα cluster between the four mouse species (**Fig S3B**). For instance, the central region of the Vα cluster (Vα gene families 3-15) is triplicated in BL6, PWD and SPRET relative to the CAST Vα cluster. This results in a Vα locus size reduction of ~0.6 Mb and approximately 70 fewer Vα genes in CAST. In a given species, sequence identity across Vα paralogues is extremely high (e.g., in BL6 *Trav11* and *Trav11D* are 100% identical on the nucleotide level; for details see [48]). For this reason, and to properly handle multiple Vα read mapping, we grouped Vα genes into their respective gene families for cross-species comparison.

We compared the mean of intra-species V/J gene segment usage across BL6, PWD, CAST and SPRET mice (**Fig. 3A**). Consistent with previous studies [33, 49], V/J segment usage within an individual typically spanned several orders of magnitude in both TCRα and TCRβ chains. For example, in BL6 TCR repertoires *Trbv12-1* and *Trbj2-7* were found in 2.37% of likely productive, in-frame, TCRβ chains, while the combination of *Trbv21* and

Trbj1-5 was only present in 0.0003% of in-frame TCRβ chains. Across species, segment usage frequencies are broadly similar, with some notable and extreme exceptions, mostly in Vβ-Segments (e.g. *Trbv13-2* is used in 7.75% of BL6 TCRβ chains compared to only 0.14% in SPRET). Consistent with previous studies [50] we observe decreased pairing of proximal Vα and distal Jα as well as distal Vα and proximal Jα segments. The only exception to this rule is CAST, where we observed significantly higher usage frequencies of the most distal Vα gene (*Trav1* and *Trav2*; chi-squared test; ** $P < 0.01$; **Fig S3D**). In laboratory mice, it has been shown that TCRα V-J recombination proceeds progressively from 3' (proximal) Vα genes towards more 5' (distal) Vα genes [51]. We therefore hypothesize, that the significantly higher usage of distal Vα genes in CAST is linked to its contracted Vα locus. As a general trend, we found that the average variance in V-segments usage frequency (var(Vβ) = 12.44, var(Vα) = 5.95) was higher compared to the average variance in J-segment usage frequency (var(Jβ) = 4.91, var(Jα) = 0.48) when compared across all species. This indicates that species-specific differences in V(D)J usage mostly arise from biases in V-segment usage rather than J-segment usage.

Further to the previous validation effort, we have also generated matching V(D)J usage profiles using the commercial 10x platform. Similar to BL6 mice, we observed excellent correlation of V-J gene usage frequencies in both approaches (Pearson correlation: PWD r = 0.89, CAST r = 0.88, SPRET r = 0.92) and only identified two V-gene segments that were not recovered in the CITR-seq dataset compared to the 10x dataset (*Trbv-31* in PWD, and *Trbv-24* in CAST/SPRET). The main difference between the platforms is that with CITR-seq we recovered on average ~160,000 T cells carrying a productive and paired TCR per experiment vs. 2,300 using the 10x platform.

Next, we performed principal component analysis (PCA) on combined TCRα and TCRβ V-J pairings across all CITR-seq replicates from each species alongside the 4 samples generated using 10x Genomics Immune Profiling, subsampling each sample to 5,000 TCRα and TCRβ chains each due to the lower throughput of the latter (**see methods**). Overall, samples are clustered strongly by species (PC1-PC3; **Fig. 3B**, **Fig S3E**) with only 6% of cross-sample variance explained by the technique (PC4, **Fig. S3E**). Mean intra-species V-J usage was highly correlated across samples for both TCRα and TCRβ (Pearson: r = 0.987 +/- 0.052 stdev vs. r = 0.991 +/- 0.044 stdev) (**Fig. 3D**). Across

samples the average V-J segment usages were more correlated for TCRα chains (r = 0.83 +/- 0.119 stdev) than TCRβ chains (r = 0.59 +/- 0.244 stdev) which is consistent with the difference in overall diversity across both chains described earlier (**Fig. 2E**). Therefore, we conclude that in the four different mouse species, V-J usage showed distinct genotype specific patterns primarily in Vβ-genes.

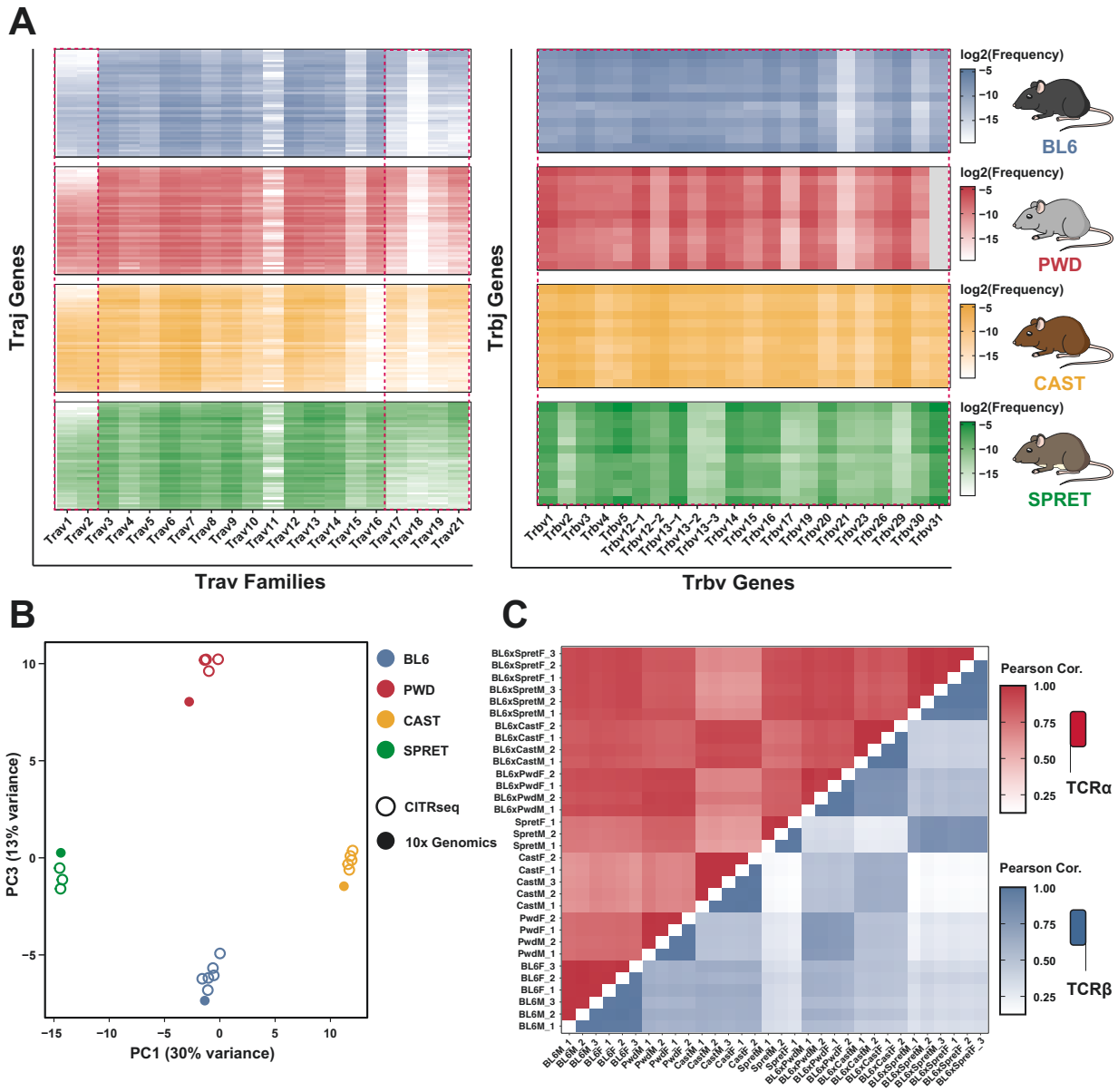


Figure 3: Species-specific V- and J-gene usage patterns in different mouse species

(A) V-J usage frequency heatmaps. Heatmap shows the frequency (log_2) of Vα-family + Jα-gene (left) and Vβ-gene + Jβ gene (right) usage of all functional TCRs in T cells across the four different mouse species (intra-species mean). Red boxes contain V-genes with one-to-one orthology in all four mouse species. J-genes are displayed in the order of their location within the locus (3' to 5'; see methods for full list). In PWD *Trbv31* is excluded due to failure of amplification during multiplex PCR.
(B) Principal component analysis (PCA) of combined Vα-Jα and Vβ-Jβ usage across all four mouse species in different samples generated using CITR-seq (empty circles) or 10x Genomics Single Cell Immune Profiling (filled circles). Samples generated using both methods cluster by genotype.
(C) Pearson correlation of inter-individual V-J gene usage in TCRα and TCRβ chains in all 32 individuals analyzed using CITR-seq in this study.

Thymic selection shapes TCR repertoire V-segment usage

Thymic selection ensures that T cells expressing TCRs with either too weak (positive selection) or to strong (negative selection) self-MHC binding properties fail to progress in the maturation process and are thus depleted from the repertoire. In effect, by collecting TCRs from peripheral (e.g., spleen-derived) T cell populations for CITR-seq, we report here the mature TCR repertoire after thymic selection. Here, we also emphasize a distinction between functional vs. non-functional TCR chains. This is because during V(D)J recombination, random insertions and deletions of nucleotides at gene-segment junctions, often introduce frameshifts or premature stop-codons. These result in transcripts representing non-functional TCRs. However, mature T cells with an in-frame (IF) TCR often still retain active transcription of an out-of-frame (OOF) TCR from its second allele that is ultimately degraded, e.g., via non-sense mediated decay [52, 53]. This presented us with an opportunity to estimate the generative usage probability of gene segments, independent from the effects of positive or negative thymic selection (see also [54]). Crucially, our use of an inbred panel of species should result in an unchanged, homozygous MHC background resulting in a consistent thymic selection regime.

Across all 32 CITR-seq sample we found 4.58×10^6 (24.4% of total transcripts) transcripts that contain frameshifts or premature stop-codons with an average per transcript UMI count of 1.78 (compared to 4.87, two-sample t-test, $P < 0.001$). To evaluate the effect of thymic selection on TCR repertoires across the different species, we compared V- and J-gene usage in OOF (pre-selection) and IF (post-selection) TCRs (**Fig. 4A**). We observed

that most V(D)J genes show similar frequencies in pre- and post-selection repertoires, summarized by normalized Shannon diversity index (nSDI, **see methods**), a measure of entropy (Vα (BL6 and PWD) as well as Jα (no significant changes) and Jβ (PWD and SPRET; **Fig. S4B**)). Again, the strong exceptions reside mostly within Vβ-genes: we observed significant differences in Vβ gene usage frequencies in all four species (**Fig. 4B**, paired t-test $P < 0.05$). The strongest absolute reduction of nSDI was observed in SPRET (-0.15) and PWD (-0.06), indicating significantly biased Vβ-segment usage in post-selection repertoires. For specific Vβ-segments, we found striking differences between pre- and post-selection repertoires, e.g., an average ~60-fold reduction in *Trbv13-2* usage frequency in SPRET post-selection repertoires (**Fig. 4C**). Notably, these extreme fold changes were mostly present in Vβ genes that showed strong cross-species frequency difference (e.g. *Trbv-2*, *Trbv12-2*, *Trbv13-2*, *Trbv17*, *Trbv21*) as shown before (**Fig. 3A**). We interpret the striking reduction in usage for these Vβ segments to be strongly suggestive of segment rejection during thymic selection.

While the most extreme differences in Vβ-segment usage tend to be species-specific, we also observed common trends shared across all four species. For instance, we observed *Trbv-2* frequencies to be consistently lower in post-selection than pre-selection repertoires across all species (log2 FC IF/OOF; BL6: -1.2, PWD -1.3, CAST -0.5, SPRET -4.4). While thymic selection acts only to remove T cells from maturation, such that the absolute number of TCRs containing a particular gene segment only decrease from pre- to post-selection repertoires, in relative terms, a given segment can be overrepresented in the final, mature repertoire through thymic selection. One such example was *Trbv-14* whose relative contribution to the TCR repertoire was higher in all post-selection repertoires across species (log2 FC IF/OOF; BL6: 1.3, PWD 1.8, CAST 1.24, SPRET 1.38).

In summary, we show that thymic selection exerts an effect on the composition of the TCR repertoire by distorting usage frequencies in all segments across all four species, but its effect is most notable in Vβ-genes, in particular in the reduction of particular Vβ-genes (e.g. *Trbv13-2*, *Trbv2*, *Trbv12-2* etc.) in PWD and SPRET, likely due to strong rejection during positive thymic selection.

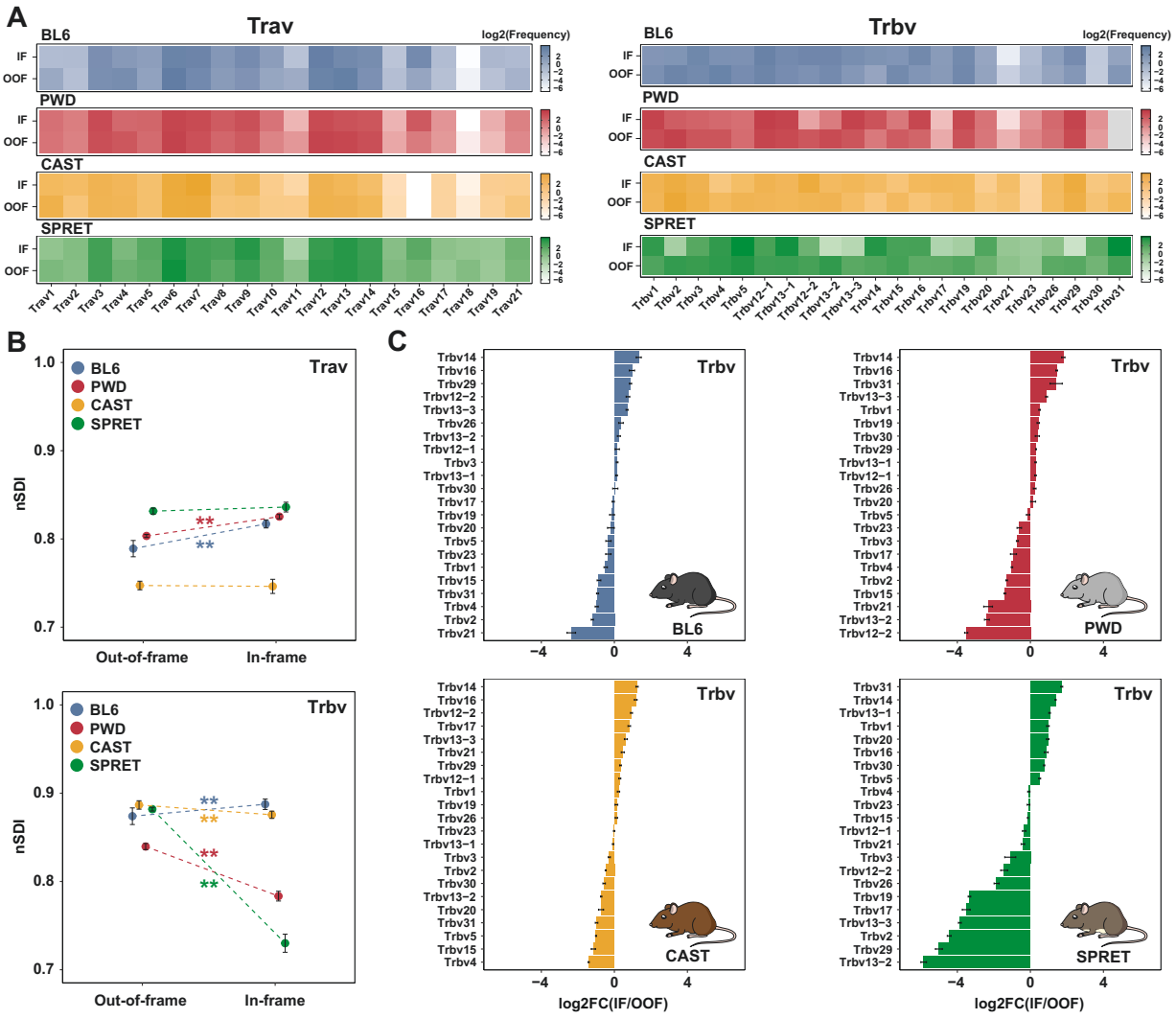


Figure 4: Thymic selection shapes V-gene usage

(A) Vα family (left) and Vβ gene (right) usage frequency (log_2) heatmaps. Heatmaps show the mean intra-species V-usage in in-frame (IF) and out-of-frame (OOF) TCRs across all T cells.

(B) Mean intra-species entropy in Vα-usage (top) and Vβ-usage (bottom) distributions calculated using the normalized Shannon diversity index (nSDI) for OOF and IF TCRs (error bars indicate the standard deviation in species replicates, significance calculated using a paired t-test, * *P-value* < 0.05).

(C) Log_2 fold-changes (log2FC) in Vβ gene usage frequencies between IF and OOF TCRs across the different mouse species (error bars indicate the standard deviation in species replicates).

Allele-specific V/J segment usage in F1 Hybrids revealed by patterns of thymic selection

In contrast to outbred individuals, assaying V(D)J gene usage in inbred mouse strains benefits from consistent thymic selection, thanks to the homozygous MHC-allele background. If so, the observed OOF-IF profile should shift in individuals carrying alternative MHC-haplotypes (see also [55, 56]). This raises further the tantalizing possibility that, depending on the actual MHC-haplotype, there may be different outcomes associated with positive vs. negative thymic selection. To test our hypothesis, we generated F1 hybrids from crosses of BL6 with each of the three other mouse species (BL6xPWD, BL6xCAST, BL6xSPRET). This gave us a powerful tool to track how the two otherwise distinct sets of species-specific V(D)J gene repertoires may be shaped by thymic selection in the respective heterozygous MHC allele state.

We first compared V(D)J usage frequencies in F1 hybrids with the respective frequencies in the parental species (**Fig. 5A** for V-genes and **Fig. S5A** for J-genes). Similar to our previous analysis, we see the most differences across Vβ-genes, in both directions: Vβ-genes can be significantly more abundant (e.g., *Trbv1*) or less abundant in F1 hybrids than in either parent (e.g., *Trbv12-1* and *Trbv12-2*; Wald-test; $P < 0.01$; **Fig. S6A-D**). To analyze the general trends across Vα, Jα, Vβ and Jβ frequencies between parental lines and their F1 hybrids, we classified their relative V(D)J gene usage frequencies into five broad categories: conserved, additive, dominant, over- and under-dominant (**Fig. 5B, see methods**). In Vα and Jα genes, 78.6% of genes only show modest frequency changes (<1%) relative to both parents. Over- and underdominance (>1% higher/lower frequency than both parents, respectively) are only seen in genes in the TCRβ chain and mostly in Vβ-genes. Notably, we observe overlaps in the identity of over-dominant (e.g., *Trbv1*) and under-dominant (e.g., *Trbv12-1*) Vβ-genes across all three hybrids. Collectively, V(D)J gene frequency changes between F1 hybrids and the parental lines are predominantly observed in the TCRβ chain.

We then calculated the nSDI for pre- and post-selection repertoires in the F1 hybrids and compared them to the previously calculated nSDI values in the parental species (Trav and Trbv **Fig. 5C** Traj and Trbj **Fig. S5B**). Across all F1 hybrids, we see significantly reduced nSDI values for Vβ-gene frequencies ($P < 0.05$; paired t-test). The increase in

unevenness between pre- and post-selection repertoires are consistently greater in F1 hybrids compared to their parental species, suggesting that thymic selection introduces stronger biases on Vβ-gene usage in F1 hybrids relative to their respective parental species. Interestingly, we observed a constant increase of nSDI values in post-selection compared to pre-selection Vα gene frequencies in F1 hybrids ($P < 0.05$ in BL6xCAST and BL6xSPRET).

Next, we took advantage of our ability to assign V(D)J genes in F1 hybrids in an allele-specific manner to identify potential biases towards usage of one parental allele. We compared the allelic ratios of V(D)J genes in pre- and post-selection repertoires (V genes: **Fig 5D, Fig S5D** and J genes: **Fig S5C**). We found significant allelic biases in Vβ-genes in post-selection repertoires that were not observed in the pre-selection repertoire. For example, in pre-selection repertoires of BL6xSPRET hybrids, ~60% of *Trbv13-2* usage was assigned to the SPRET allele, whereas in post-selection repertoires this rate dropped to ~1%. Therefore, while in BL6xSPRET hybrids the *Trbv13-2* allele was frequently recombined during V(D)J recombination, it was almost completely rejected during thymic selection. The almost exclusive selection of one parental allele in a heterozygous MHC haplotype and a common *trans*-environment, provides strong evidence that this selection process is primarily determined by genetically encoded polymorphisms in the underlying Vβ gene. Similarly, we saw that while Traj35 pre-selection frequencies are balanced between parental alleles (Percent of BL6 alleles: PWD 49%, CAST 44%, SPRET 50%), the BL6 allele was substantially less frequent in post-selection repertoires (Percent of BL6 alleles: PWD 25%, CAST 26%, SPRET 25%; **Fig S5C**).

Apart from these strong exceptions, allelic bias is strongly correlated in pre- and post-selection repertoires for most V(D)J genes (see Pearson correlation in **Fig. 5D** and **S5C**). Genes that show strong frequency differences between both parental species (*Trav16* and *Trbv21* in BL6/CAST, *Trav18* in BL6/SPRET or *Trbv17* in BL6/PWD) often show strong F1 allelic bias towards usage of the respective parental allele that had a higher frequency in the pure contrast (**Fig. 5D**). We therefore conclude that the (generative) pre-selection biases observed between species are primarily controlled by linked factors acting in *cis*, e.g., polymorphisms in the RSS sequences that influence the recombination likelihood of a particular gene during V(D)J recombination.

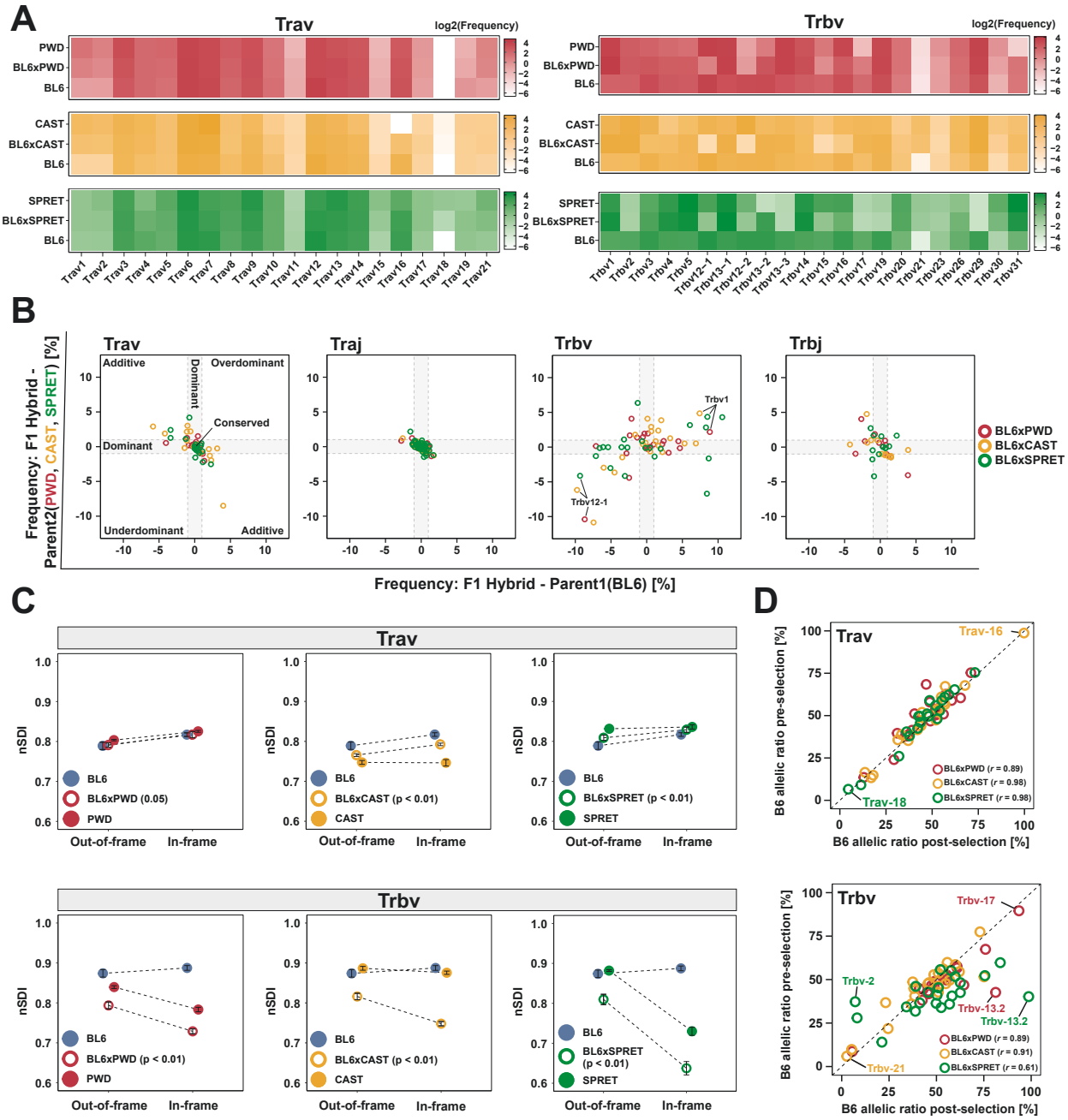


Figure 5: V-gene usage is more restricted in F1 hybrids and shows allele specific usage biases

(A) Vα family (left) and Vβ gene (right) usage frequency (log_2) heatmaps of in-frame TCRs in F1 hybrids and their respective parental species.

(B) Relative frequency changes of V(D)J gene usage in F1 hybrids and the respective parental species (x-axis: F1 hybrid – BL6 and y-axis: F1 hybrid – PWD, CAST or SPRET) categorized into mode of inheritance. Conserved (center), dominant (grey area), additive (top left and bottom right quadrant), under-dominant (bottom left quadrant) and over-dominant (top right quadrant; see methods). Each circle represents a Vα-family, Jα-gene, Vβ-gene or Jβ-gene.

(C) Comparison of entropy of V-usage distribution in F1 hybrids and the respective parental species calculated using the normalized Shannon diversity index (nSDI) for OOF (left) and IF (right) TCRs (error bars indicate the standard deviation in species replicates, significance tested for F1 hybrid IF vs OOF contrast using paired t-tests, *P-value* < 0.05).

(D) Analysis of biased V gene allele usage in F1 hybrids. Plots show the percentage of BL6 Vα family alleles and Vβ gene alleles in post- (x-axis) and pre-selection (y-axis) TCRs. Each circle represents a Vα-family (top) or Vβ-gene (bottom). Pearson-correlation was calculated for post- and pre-selection V gene usage.

Composition and diversity of the paired CDR3αβ repertoire depends on an individual's genotype

Next, we addressed CDR3 diversity 32 samples, representing seven different genotypes. We reasoned that given the cumulative bias in gene segment usage, availability, as well as genetic differences, we should observe distinct CDR3 amino acid motif repertoires. To test this hypothesis, we first compared the CDR3α, CDR3β and CDR3αβ diversity in a set of 100,000 T cells sampled randomly from each individual (**Fig. 6A**). We found that, across all comparisons, F1 hybrids show an increased number of unique CDR3 sequences relative to their respective parental species. In single CDR3α motifs, we see significant differences in diversity within the parental species, which is in line with the observed differences in locus structure of the TCRα loci across these mice (e.g., ~50 fewer Vα segments in CAST compared to all other species; pairwise t-test; ** = $P < 0.01$; * = $P < 0.05$). These parental diversity differences are not recapitulated in the F1 hybrids. Instead, the absolute diversity increases with the increasing evolutionary divergence of the parental species.

A strikingly different picture emerged for the CDR3β motifs. Importantly, V(D)J segments in the TCRβ locus follow a strict one-to-one orthology across all parental species. Accordingly, single-chain CDR3β diversity showed little variation across the parental species. In contrast, F1 hybrids showed much greater variation and generally display greater TCR diversity than observed in the repertoire of either parent (pairwise t-test; ** = $P < 0.01$; * = $P < 0.05$). A possible explanation for this is the strictly restricted selection of Vβ segments during thymic selection in the TCRβ chain. We observe remarkable diversity of paired CDR3αβ across all genotypes, with an average of 98.2% of all motifs being observed only once in each set of 100,000 motifs. Notably, among F1 hybrids, the lowest diversity is observed in BL6xSPRET hybrids despite the highest evolutionary divergence in the respective parental species.

Due to random insertions and deletions at the segment junction sites, the highest amino acid diversity within CDR3 motifs is observed in the central region of the peptide chains (**Fig**. **S7A**). It has been shown that this central region overlaps the region of highest antigen proximity in TCR-MHC complexes (position 107-115 according to IMGT nomenclature [57]) and therefore contributes most to the antigen specificity of the underlying TCR [6]. Further, the same study also provided evidence that antigen-specificity is defined by specificity-groups of similar amino acid motifs within TCRs. With this in mind, we analyzed germline-encoded differences in the central motifs across all mice. Because the same antigen might be recognized by several similar TCRs rather than just one CDR3αβ motif we first generated amino acid 4mers from both CDR3 motifs of TCRs of individual cells (**Fig 6B**). We then identified a list of 1,201,646 common 4mers across all genotypes (**see methods**). Next, we performed PCA analysis based on the abundance of all 4mer pairs across all 32 individuals. We see that 4mer pairs are strictly clustered according to the underlying genotype of each sample (**Fig. 6B**). This pattern was also observed in the corresponding analysis on single-chain derived unpaired 4mers (**Fig S7B**).

Comparison of TCR repertoires often involves the analysis of shared “public” CDR3 motifs. Typically, this type of analysis addresses motif sharing within single chains across repertoires. While these comparisons might provide information on the generative probability of distinct single-chain CDR3 motifs across individuals, the missing CDR3 motif in the second chain makes it challenging to identify potential shared TCR responses to antigens. Here, we utilized CITR-seq’s large set of more than 5 million paired CDR3 motifs to analyze motif sharing across all individuals. In total, we identified 25,894 (~0.5% of all motifs) paired motifs with identical amino acid sequence, observed in different individuals. Across single chains, sharing of identical amino acid sequences was more common with 264,088 shared CDR3α (~36.7% of all unique motifs) and 469,827 shared CDR3β (~ 27.2% of all unique motifs) motifs observed in at least two individuals. Notably, we found 1,696 CDR3α and 644 CDR3β amino acid motifs that were observed in all 32 individuals, while identical CDR3αβ pairs were at most observed in 12 individuals (**Fig. S7C**)

To test whether the extent of paired CDR3αβ motif sharing was higher than expected by chance, we shuffled the α- and β-chains within each individual and then re-calculated the count of shared motifs. We saw that the observed sharing count is about 4-fold higher than the mean across 100 permutations of αβ-chains shuffled samples (mean: 6182 shared CDR3αβ motifs across shuffled pairs, permutation test *P-value* < 0.01). Next, we analyzed whether the extent of sharing in CDR3αβ motifs is dependent on the underlying genotype of each sample. To account for the variance in sample size across all samples, we calculated the Jaccard Index of repertoire sharing using paired CDR3αβ motifs (**Fig. 6C, see methods**). We observed that motif sharing is significantly higher across samples of identical genotypes (56.7% of all shared motifs), compared to individuals with partially shared genotypes (F1 hybrid samples, 32.0% of all shared motifs) and especially in contrast to completely unrelated individuals (11.3% of all shared motifs; Wilcoxon rank sum test *P-value* < 0.01) (**Fig. 6D**). While we caution here that our use of inbred individuals may differ from the usual comparison contexts with CDR3 motifs, nevertheless, the extent of sharing across fully unrelated individuals led us to conclude that an individual's genotype contributes significantly to the final TCR repertoire. Additionally, public TCR responses are far more likely to be observed across related individuals than unrelated individuals.

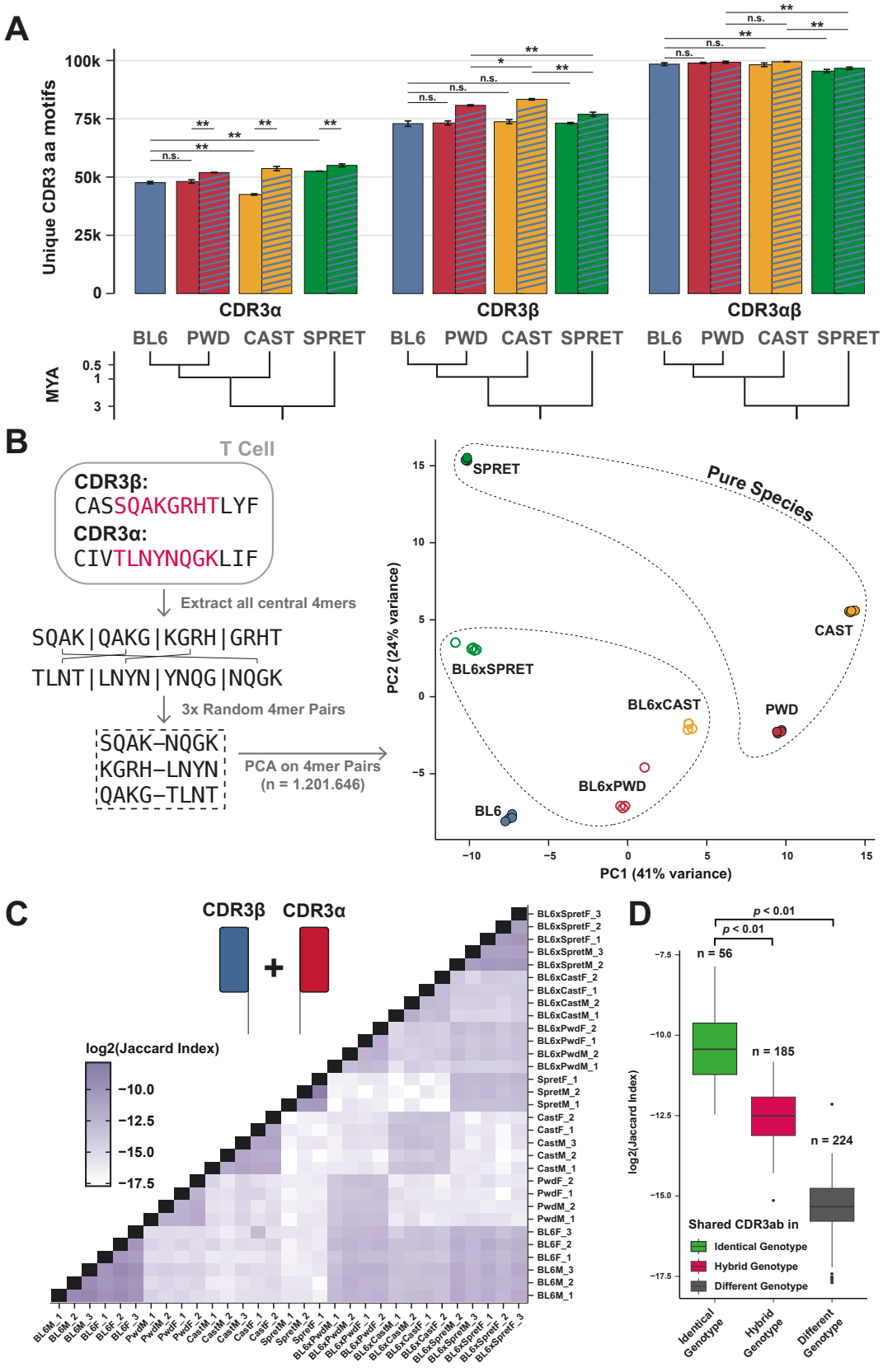

A
Unique CDR3 aa motifs
100k
75k
50k
25k
0
n.s.
**
*
CDR3α
CDR3β
CDR3αβ
BL6 PWD CAST SPRET
MYA
0.5
1
3
B
T Cell
CDR3β:
CASSQAKGRHTLYF
CDR3α:
CIVTLNYNQGKLIF
Extract all central 4mers
SQAK|QAKG|KGRH|GRHT
TLNT|LNYN|YNQG|NQGK
3x Random 4mer Pairs
SQAK–NQGK
KGRH–LNYN
QAKG–TLNT
PCA on 4mer Pairs
(n = 1.201.646)
SPRET
Pure Species
CAST
BL6xSPRET
BL6xCAST
PWD
BL6xPWD
BL6
PC2 (24% variance)
PC1 (41% variance)
C
CDR3β CDR3α
log2(Jaccard Index)
−10.0
−12.5
−15.0
−17.5
D
p < 0.01
p < 0.01
n = 56
n = 185
n = 224
log2(Jaccard Index)
Shared CDR3ab in
Identical Genotype
Hybrid Genotype
Different Genotype

Figure 6: CDR3 motif diversity and sharing depends on an individual's genotype

(A) Mean count of unique CDR3α (left) CDR3β (middle) and paired CDR3αβ (right) amino acid motifs in a set of 100,000 randomly sampled TCRs from each individual grouped by genotype (error bars indicate the standard deviation in species replicates, significance calculated by pairwise t-tests with * *P-value* < 0.05 and ** *P-value* <0.01). Single-color bars represent the parental species, diagonally striped bars represent the respective F1 hybrids. The phylogenetic tree (bottom) shows the evolutionary divergence of parental species.

(B) Analysis of paired 4mers extracted CDR3αβ motifs. For each paired CDR3αβ amino acids sequence all possible 4mers were extracted and subsequently, 3 random 4mer pairs (one from the CDR3α and one from the CDR3β sequence) were generated (left panel). The combined filtered (see methods) count matrix of 4mer pairs from all 32 individuals was then used for PCA analysis. Samples cluster based on the sample genotype.

(C) Overlap of paired CDR3αβ amino acid motifs between all 32 CITR-seq samples calculated using the Jaccard index (log_2; see methods)

(D) Based on overlap of genotypes all samples were grouped into identical genotype (within species, e.g., BL6F_1 and BL6M_1), hybrid genotype (50 % identical genotype, e.g., CAST and BL6xCAST) and different genotype (completely unrelated individuals, e.g. PWD and SPRET). Boxplot shows the calculated Jaccard index values (log_2) in each respective group (significance tested using Wilcoxon rank-sum test, *P-value* < 0.01).

Discussion

Production and maintenance of large and diverse repertoire of TCRs is crucial for a functioning adaptive immune system. For decades researchers have now accumulated insights into the generative process, the size and overlap, as well as associations to disease states of TCR repertoires. High-throughput sequencing technologies have reached sufficient sensitivity and throughput to capture reasonable portions of an individual's TCR repertoire. Yet, they still suffer from severe limitations in the face of the extreme diversity of TCR repertoires. To date, arguably the most limiting of these factors is the requirement for single-cell resolution to link both TCR chains of the heterodimeric αβ receptor to the T cell of origin. With few (mostly non-commercial) exceptions, single-cell TCR sequencing methods suffer from low-throughput (10^3-10^5 T cells) and high cost (reviewed here [58]). Pit against the vast TCR repertoire diversity, especially in naïve repertoires, those technologies often capture only a tiny fraction of an individual's repertoire.

In this study we present CITR-seq, a high-throughput low-cost single-cell TCR sequencing method that overcomes many of these limitations. We use CITR-seq to generate TCR repertoires of four evolutionary divergent inbred mouse species and t their respective F1 hybrids, covering more than 9 million T cells with 76% successful αβ-pairing rate.

We first identified large differences in V(D)J gene usage across the different mouse species, with very high within-species consistency for both TCR chains. While the arrangement and number of genes in the TCRβ locus are conserved across all species, the TCRα locus has undergone complex rearrangements leading to triplications and inversions of Vα gene clusters. As a result, the number as well as their relative distance to Jα genes varies substantially between Vα genes of the different mouse species. We observed that at the TCRα locus in CAST, in which the Vα locus was contracted by 0.6 Mb, the distal Vα genes showed significantly higher segment usage compared to the other species. Considering the progressive 3' to 5' recombination of V-J segments [59], we interpret this as evidence for a direct relation of gene segment locus size and chromosomal position dependent usage frequency.

Due to the very conserved arrangement of TCRβ genes, the tightly enforced allelic exclusion as well as prevention of continuous rearrangements, the relative position of genes should contribute less to biases in the TCRβ gene usage across species. Nevertheless, we see that the relative fold-changes in gene usage of Vβ genes can be extreme, with up to 60-fold difference between different mouse species.

We show that many of those extreme gene usage differences are introduced during thymic selection by comparing pre- vs post-selection repertoires. We use the nSDI of segment usage to demonstrate that thymic selection primarily acts on Vβ segments and show that their generative frequency immediately after V(D)J recombination is more similar across different mouse species than the actual usage frequencies observed in mature and selected TCR repertoires. Critically, we observed that many of the Vβ genes that are rejected during thymic selection, contribute identical amino acids to CDR3 motifs compared to other Vβ genes that do not significantly change in frequency in pre- vs. post-selection repertoires. Thus, we hypothesize, that the rejection of those Vβ genes is unlikely to be enforced during negative selection as a consequence of strong affinity to a self-MHC complex. Rather, it is reflective of their particular germline-encoded ability to bind MHCs evaluated during positive selection. This hypothesis is further supported by the fact that we did not observe categorical rejection of J-segments, that are though to mostly contribute to the antigen specificity of a TCR rather than its ability to bind to MHCs.

Further experiments, where both Vβ and MHC components can be experimentally controlled, may be able to shed light on the mechanism underlying our observation.

We also used F1 hybrids of inbred mouse species, as a powerful tool to evaluate the thymic selection of TCRs in a defined heterozygous MHC haplotype. In those hybrids, two sets of V(D)J genes are exposed to a common *trans*-environment that subject both to a common positive and negative selection regime. As a general trend, we see that most V(D)J genes show conserved usage frequencies relative to the parental species, or alternatively, in the case of substantial differences between the parental species, exhibit intermediate (*additive*) gene usage frequencies. These general patterns are far less pronounced for Vβ gene usage frequencies of F1 hybrids. We see that the selection against particular Vβ genes mostly resembles the patterns seen in the parents, with additional rejections of particular genes that were frequent in both parents. By utilizing our species-specific V(D)J references, we were able to disentangle the usage frequencies of particular alleles in F1 hybrids. We provide examples of Vβ-genes with balanced allelic ratios in pre-selection repertoires and striking allelic biases in post-selection repertoires. The nearly mono-allelic usage of particular Vβ-genes as a consequence of thymic selection in a defined heterozygous MHC allele state in F1 hybrids, provides strong evidence that the rejection of particular Vβ-gene alleles is based on genetically encoded polymorphisms. To the best of our knowledge, such extreme cases of allele-specific Vβ-genes selection have not been described before. This finding has important implications for the ongoing debate about whether binding to MHCs is an inherent and germline-encoded feature of TCRs that progressively co-evolves, or alternatively, MHC restriction of TCRs is enforced by TCR co-receptor signaling involved in TCR-MHC complex formation. Due to the common *trans*-environment during thymic selection of TCRs, the strong allelic biases of particular Vβ genes can hardly be explained by co-receptor signaling and thus should reflect the inherent ability of particular Vβ gene alleles to bind MHCs originated from heterozygous alleles in the F1 hybrids. Consequently, we hypothesize that TCR-MHC binding is a co-evolutionary process mediated by changes in amino acid sequences of V gene regions and MHC alleles that facilitate complex formation. In this context, the highly variable germline-encoded CDR1 and CDR2 regions of TCR V-genes have been shown to be crucial for altering TCR-MHC binding strength.

A secondary consequence of the co-evolution of TCR-MHC binding would likely be the increased rate of TCRs that exhibit insufficient or overly strong affinity to MHCs in hybrids between highly divergent parents. Indeed, as shown in this study, thymic selection had the strongest effect on Vβ genes in BL6xSPRET individuals in which the respective parental individuals had the highest degree of evolutionary divergence.
We further show that biases that are consistent in pre- and post-selection repertoires mostly reflect selection independent gene usage frequency differences observed in the parents. For instance, TCR consisting of *Trbv21* are extremely rare in BL6 (0.03% of TCRs) but much more frequent in CAST (3.0% TCRs) with minor differences in pre- and post-selection frequency. In BL6xCAST F1 hybrids, 2.2% of all TCRs consist of *Trbv21* with an allelic ratio of 97.3% of CAST alleles and only 2.7% BL6 alleles. Therefore, gene segment frequency biases are mediated through *cis*-effects in the absence of any additional biases introduced by thymic selection. For instance, polymorphisms in the RSS in between V(D)J genes could bias the recombination efficiency of particular gene segments.

What are the consequences of the observed frequency and selection biases across the different mouse species for the total diversity within the TCR repertoires? To answer this question, we evaluated CDR3 motif diversity in single chains as well as paired TCRs. CDR3α diversity varies most across the pure species, which is likely caused by the severe rearrangements and consequently different number of functional gene segments in the Vα cluster. We generally observe minor frequency changes of Vα families in pre- and post-selection repertoires, indicating that those gene segments are subject to less stringent thymic selection. As a consequence, F1 hybrids can make full use of both parental sets of V(D)J genes, which likely leads to the correlation of increased CDR3α diversity with increased evolutionary divergence of parental species. Here, we note that grouping of Vα genes by their respective families might mask the rejection of particular genes during thymic selection. While this potentially impacts the gene usage frequency differences across the species, it does not bias the comparison of total CDR3α diversity.
The single-chain CDR3β motif diversity is extremely similar across pure species, which is in line with the one-to-one orthology of V(D)J genes in the TCRβ locus. In contrast to

this, we see substantial differences in CDR3β diversity in the hybrids. Based on the observed impact of thymic selection on Vβ genes, we hypothesize that the observed differences in CDR3β diversity in F1 hybrids result from a trade-off between the diversity of parental V(D)J gene sets and the increased likelihood of gene segment rejection during thymic selection, which should correlate with the increasing evolutionary divergence of parental species. While in this study, thymic selection is evaluated in a fixed and genotype-specific MHC-haplotype set-up, it has been shown that increased intra-individual MHC diversity is associated with increased rates of T cell depletion during thymic selection [60, 61]. Given that HLA allele frequencies vary substantially across human populations [62], we assume that the general trends observed in this study would therefore also apply in the context of CDR3 diversity evaluation in evolutionary divergent outbreed populations exhibiting diverse MHC haplotypes.

To the best of our knowledge, the present study analyzes the largest set of paired αβ-TCRs to date. Especially in the context of TCR repertoire analysis of antigen in-experienced naïve T cells we benefit greatly from the scale of our dataset. Sharing of identical CDR3αβ motifs is rare but about 4-fold higher than expect by chance. Additionally, shared motifs are found at significantly higher rates in related individuals compared to unrelated individuals. Importantly, the increased sharing rate of paired CDR3αβ motifs analyzed in this study is not limited to the comparison of 100% identical motifs. Because similar CDR3αβ motifs might recognize identical antigens and similar antigens might be recognized by a range of similar CDR3αβ motifs we used a kmer-based approach to emphasize the similarity of paired CDR3αβ motifs in species of identical genotypes. We showed that 4mers originated from the central region of paired CDR3αβ motifs exhibit remarkably similar frequencies in species with identical genotypes relative to unrelated individuals. We therefore conclude that the combined effects of differences in TCR locus structure, V(D)J recombination frequencies and biases introduced by thymic selection, collectively shape the TCR repertoire in a genotype-specific manner.
This also has important implications for our understanding of public TCR motifs with potential disease associations. The number of shared CDR3 motifs in individuals with diverse MHC haplotypes is representative of those TCRs, that are selected by the specific

set of MHCs in the sampled individuals. Public CDR3 motifs should therefore always be cataloged in the specific MHC haplotype context they have been observed in to allow for the comparison of such public motifs across different studies. Additionally, public TCR responses are often evaluated in common disease context, such as cytomegalovirus (CMV) and Epstein-Barr virus (EBV) [63-65]. Since large parts of human populations are persistently infected by those pathogens, a broad range of MHC haplotypes should have evolved to effectively present EBV- and CMV-derived peptides. Consequently, EBV- and CMV-associated CDR3 motifs might be more public compared to CDR3 motifs that specifically recognize less frequent pathogenic peptides.

Immune receptor diversity is one of the most characteristic and important features of adaptive immunity. While the generation of diversity is in large parts driven by stochastic events, the present study highlights important genetic contributions to TCR diversity. We show that the number of functional V(D)J segments, their *cis*-regulated recombination frequency as well as MHC haplotype dependent thymic selection, collectively generates TCR repertoires that are significantly more similar within than across genotypes.

Methods

Mice

All mice were housed in the animal facility of the Friedrich-Miescher Laboratory of the Max-Planck Society. Experiments were performed under license issued by the local competent authority (EB 01/21 M). Spleens were collected from mice aged 9-11 weeks. The following mouse strains were used in the experiments: C57BL/6J (The Jackson Laboratory, Strain #: 000664), CAST/EiJ (The Jackson Laboratory, Strain #: 000928), SPRET/EiJ (The Jackson Laboratory, Strain #: 001146), PWD/PhJ (The Jackson Laboratory, Strain #: 004660) as well as their respective F1 hybrids (C57BL/6J x SPRET/EiJ/CAST/EiJ/ PWD/PhJ). Male and female mice of all strains were used.

Isolation of CD8a$^+$ T-cells

Spleens of euthanized mice were collected and placed on a 40µm cell-strainer. Spleens were then pressed through the strainer using the backside of a syringe plunger. After thorough rising of the cell-strainer using ice-cold PBS, the flow-through was centrifuged at 400xg 4°C for 10 minutes in a swing-bucket centrifuge. Afterwards, supernatant was carefully discarded, and the cell pellet was resuspended in 1ml ice-cold PBS + 2% FBS. Isolation of CD8a$^+$ T-cells was then done using the "Dynabeads™ FlowComp™ Mouse CD8 Kit" (Invitrogen, 11462D) according to the manufacturer's instructions. Pre-enriched cells were then stained using anti-CD4 BV510 (Bio Legend, 100553) and anti-CD8 PerCP-Cy5.5 (Bio Legend, 155013) in 500µl PBS + 2% FBS for 15 minutes on ice. Afterwards, cells were centrifuged at 400xg 4°C for 5 minutes. Supernatant was discarded and cell pellet was resuspended in 500µl ice-cold PBS + 2% FBS. This washing step was repeated once before final resuspension in 1 ml ice-cold PBS + 2% FBS. Cells were then further purified by fluorescence activated cell sorting (Fig. S1A). Depending on the size of the spleen (approx. 20mg in SPRET and up to 100mg in BL6) between $1x10^6$ and $5x10^6$ CD8+ T-cells were isolated from each spleen. Isolated T-cells were immediately transferred to prepared tissue culture dishes or used as primary cells for CITR-seq experiments.

Tissue Culture

Tissue culture of isolated CD8$^+$ T-cells was done as described by Lewis et al. [66]. Briefly, 6-well plates were coated with 0.5µg/ml anti-CD3 and 5µg/ml anti-CD28 in 3ml PBS at 4°C overnight. Before seeding the isolated CD8$^+$ T-cells, plates were washed twice with PBS. Cells were cultured in RPMI 1640 medium (ThermoFisher, 11875093) supplemented with 10% FBS, 1% GlutaMAX (ThermoFisher, 35050061), 1% penicillin/streptomycin (ThermoFisher, 15140122), 0.1% 2-mercaptoethanol (ThermoFisher, 21985023) and 0.1% human recombinant insulin (ThermoFisher, 12585014) at 37°C, 5% CO_2. After 20 hours cells were washed once with culture medium and then carefully detached from plate by repeatedly flushing the plates with a P1000 pipette. The cell suspension was then centrifuged at 400xg, RT for 5 minutes. Afterwards cell pellet was resuspended in 1ml PBS.

CITR-seq protocol

Oligonucleotides for barcoding

Two rounds of barcoding, each with 192 unique DNA barcodes are performed in CITR-seq. To prepare the barcoding plates in each well of two 96-well plates one unique round 1 top-stand oligo and one corresponding round 1 bottom-strand oligo were diluted in 10µl annealing buffer (10mM Tris pH 8, 50mM NaCl and 1mM EDTA). Top-strand round 1 oligos are partially complementary to the 5` overhang of the RT primers and anneal to the complementary sequence of the round 1 bottom-strand including the 7bp barcode sequence. Round 1 bottom-strand oligos contain a common 3bp 5` phosphorylated linker overhangs (“TCT”). The same procedure was repeated for two 96-well round 2 barcoding plates. Round 2 top-strand oligos contain a 3`-linker sequence (“AGA”) complementary to the 5’ linker sequence of round 1 oligos. Further, it contains another unique 7bp DNA barcode and the standard Illumina TrueSeq i7 sequencing adapter (Illumina, see document: 1000000002694). Round 2 bottom-strand oligo is complementary to its respective round 2 top-strand mate but lacks the 3bp linker sequence.

Oligos are used at the following concentrations: For each well of round 1 plates: µM of round 1 bottom-strand and µM of round 1 top-strand. For round 2 plates: µM of round 2 bottom-strand and µM of round 2 top-strand. Prior to each experiment round 1 and round 2 oligo plates are annealed in a PCR machine by heating plates to 90°C and then decreasing the temperature by 1°C every 30 seconds until room temperature is reached.

Oligonucleotides for reverse transcription

To increase the barcoding space further, barcoded RT-primers are used. Eight pairs of RT-primers targeting the constant region of the TCR alpha and TCR beta locus were designed with a 4bp barcode and a 10bp UMI as well as a phosphorylated 5’ overhang complementary to the overhang of the round 1 top-strand barcoding oligo.

TCR-V-segment primer pool for multiplex PCR

Primers were initially designed by alignment of annotated C57BL/6J cDNA sequence (IMGT database) belonging to the same TCR-V-segment family. For each family 1-5

primers (depending on number and sequence similarity of TCR-V-segment families) with similar annealing temperature (+/- 1°C), length and G/C content were designed (see supplementary table X). Subsequently, C57BL/6J TCRα and TCRβ loci were aligned to the corresponding genomic sequence in the genomes of CAST/Ei, PWK/PhJ (evolutionarily closest publicly available genome compared to the used PWD/PhJ mouse strain) and SPRET/EiJ (genome data available as part of the Mouse Genome Project from Sanger Institute). Candidate primers were then BLAT searched against the aligned genomes to rule out the presence of SNPs in the primer binding region across all strains. All candidate primers were individually tested to exclusively amplify the corresponding V-segment(s) in reverse transcription reactions using RNA isolated from C57BL/6J $CD8a^+$ T cells.

The final set of TCR-V-Segment primers consists of 58 individual primers (19 Vβ and 39Vα primers). Additional to the V-segment specific 3` end of the primer, each primer also contains a common 5` sequence used as target in the index-PCR. All V-segment primers were pooled at an equimolar ratio with a final concentration of 100 µM (1.72 µM of each primer). The primer pool was prepared once, and aliquots were frozen until used in an experiment to prevent biases introduced by varying primer pools across all experiments.

Cell fixation

After cell collection from tissue culture plates, 1ml of cell-suspension in PBS was added to 2.8ml of ice-cold PBS with 200µl of 16% PFA (ThermoFisher, 28908), for a final concentration of 0.8% PFA. After 10 minutes of incubation on ice 150µl 10% Triton-X was added to permeabilize cells and incubation on ice was continued for another 3 minutes. Cells were then centrifuged at 400xg 4°C for 5 minutes. Supernatant was discarded and the cell pellet was resuspended in 500µl 0.6M Tris-HCL pH8. Afterwards, 500µl of wash-buffer (PBS + 2% FBS and 0.4U/µl RNAseInhibitor (JenaBioscience, PCR-392L)) was added and cells were centrifuged at 400xg 4°C for 5 minutes. Washing was repeated once with 1ml wash-buffer before cells were counted and the concentration was adjusted to 50.000 cells/ml with wash-buffer.

Reverse transcription

10µl of fixed cells (~50.000 cells) were added to each of 8 tubes of a prepared PCR-strip containing 1µl 10µM barcoded TCRalpha constant region RT-primer, 1µl 10µM barcoded TCRbeta constant region RT-primer, 7.5µl NEB TS Buffer (NEB, B0466SVIAL) and 2µl 10mM dNTPs (ThermoFisher, R0181). TCRalpha and TCRbeta RT-primers within each tube share the same tube-specific 4bp barcode. The number of reverse transcription reactions can be scaled up easily by increasing the number of prepared PCR strips. Typically, two PCRs strips for a total of 16 reverse transcription reactions were prepared resulting in a final cell count of ~600.000 after barcoding (during the barcoding procedure about 25% of cells are lost due to repeated transferring and pooling of cells). Cells were then heated to 55°C for 5 minutes and rapidly cooled down to 4°C to allow pre-annealing of the RT-oligos to their target mRNAs. Afterwards, 6.3µl water, 1.5µl Maxima H Minus Reverse Transcriptase (ThermoFisher, EP0751) and 1ml RNAseInhibitor (JenaBioscience, PCR-392L) was added to each reaction for a final reaction volume of 30µl. Reverse transcription was carried out under the following conditions: 50°C for 10 minutes followed by 3 cycles of (8°C for 12 s, 15°C for 45 s, 20°C for 45 s, 30°C for 30 s 42°C for 2 minutes and 50°C for 3 minutes) and a final incubation at 50°C for 10 minutes. After reverse transcription cells were centrifuged at 400xg 4°C for 5 minutes. Supernatant was carefully discarded without disturbing the cell pellet. Cells were then resuspended in 50µl wash-buffer per tube and pooled in one 5ml tube and washing was repeated once.

Barcode ligation

All tubes used for pooling and washing of cells were coated with PBS +2% FBS to prevent cells from sticking to the plastic. Cells were resuspended in 2ml ligation buffer 1 (1460 µl water, 400 µl 10x T4 DNA ligase reaction buffer (NEB, B0202SVIAL), 100 µl T4 DNA ligase (NEB, M0202LVIAL), and 40 µl 10% Tween-20). 10µl of cell suspension was pipetted to each well of the two 96-well round 1 barcoding plates, taking care to not touch the liquid at the bottom of the plate. Plates were sealed with adhesive seals (ThermoFisher, AB0558) and incubated on a shaker for 40 minutes at room temperature. Afterwards, 3.5µl blocking oligo solution (20µM blocking oligo in water) was added to each well of both round 1 barcoding plates and incubation was continued for additional 20

minutes. The blocking oligo anneals to un-ligated round 1 top-strand oligos to prevent undesired ligations during the first cell pooling. Using a multichannel pipette, cells from both round 1 barcoding plates were pooled into a reservoir and then transferred to a 5 ml tube. Afterwards cells were centrifuged at 750xg 4°C for 3 minutes, supernatant was discarded, and cells were resuspended in 5ml of ligation buffer 2 (2260 µl water, 700 µl T4 DNA ligase reaction buffer, 100 µl T4 DNA ligase, 1900 µl annealing buffer and 40 µl 10% Tween-20). 25µl of cell suspension was pipetted into each well of the two 96-well round 2 barcoding plates, again without touching the liquid at the bottom of the wells. Plates were sealed and incubated for 40 minutes on a shaker at room temperature. Cells were then pooled as described before, centrifuged at 750xg 4°C for 3 minutes and resuspend in 200µl wash buffer. 1x DAPI (ThermoFisher, D1306) was added, and cells were counted on the Evos Countess II. The concentration of cells was adjusted to $2x10^6$ cells/ml and 5 µl of cell suspension was transferred to separate tubes of PCR-strips for the generation of sub-libraries. The number of cells in each sub-library determines the expected number of barcode collisions in each sub-library. The number of collisions can be calculated with the formular used in the birthday problem. Here the total number of barcodes B is 294.912 (8 reverse transcription barcodes * 192 round 1 * 192 round 2 ligation barcodes) with a cell count of N = 10.000 cells per sub-library. The number of expected barcode collisions therefore is:

$$10000 - 294912 + 294912\left(\frac{294912 - 1}{294912}\right)^{10000} = 167$$

With 167 barcode collisions the expected collision rate is ~1.67% in each sub-library.

Reverse Crosslinking

8 µl reverse crosslinking buffer (1% SDS, 100mM Tris-HCl pH8 and 100mM NaCl), 2 µl Proteinase K (Qiagen, RP107B-1) and 5 µl water was added to each tube with 5 µl sub-library for a final volume of 20 µl per reaction. Reverse crosslinking was done at 62°C for 2 hours on a shaker followed by a final incubation at 95°C for 15 minutes to inactivate Proteinase K. Afterwards, 12µl 10% Tween-20 was added to each sub-library to quench SDS before PCR.

cDNA library preparation

After reverse crosslinking and SDS quenching 48µl multiplex-PCR mix (23 µl water, 16 µl 5x Q5 reaction buffer, 3.2 µl TrueSeq-i7-long primer, 3 µl 10mM dNTPs, 2 µl 100 µM TCR-V-Segment primer pool and 0.8 µl Q5 DNA polymerase) was directly added to each sub-library for a final PCR reaction volume of 80µl. PCR was done using the following parameters: 98°C 2min, then 10 cycles of (98°C 20 s, 63°C 30 s, 72°C 2 minutes) and a final incubation at 72°C for 5 minutes. After PCR amplified cDNA was purified by bead clean-up using custom size-selection beads at a ratio of 1.2x beads to PCR reaction (100 µl beads) to get rid of excess primers from the multiplex PCR. During this clean-up it is important to not cross-contaminate different sub-libraries as they have not yet received their sub-library specific index.

14.5 µl index-PCR mix (10 µl 5x Q5 reaction buffer, 2 µl TrueSeq-i7-long primer, 2 µl 10mM dNTPs and 0.5 µl Q5 DNA polymerase) was added to each sub-library. Afterwards, 2.5µl of a unique 10µM Nextera N5xx primer was added to each sub-library for a final reaction volume of 50 µl. Index PCR was done using the following parameters: 98°C 2min, then 12 cycles of (98°C 20 s, 63°C 30 s, 72°C 2 minutes) and a final incubation at 72°C for 5 minutes. After index PCR sub-libraries were purified using 1.2x size-selection beads as described above. cDNA concentration of each sub-library was measured, and sub-libraries were then pooled at an equimolar ratio. Before freezing the pooled libraries until sequencing they were quantified using the Qubit HS dsDNA Quantification Kit and run on the Agilent 2100 bioanalyzer with a High Sensitivity DNA kit.

DNA size selection with custom beads

To prepare custom DNA size-selection beads, 750 µl of SPRIselect (Beckman Coulter, B23318) were transferred to a 1.5 ml tube and placed on a magnetic stand. Supernatant was discarded and beads were washed once with 1 ml Tris-HCl pH 8. Beads were then resuspended in 50 ml bead buffer (22 mM PEG-8000, 2.5 M NaCl, 10mM Tris HCl pH 8, 1 mM EDTA in water).

In general, size selection beads are added to the solution containing DNA at a defined ratio to bind DNA of a specific length (e.g., 1.2x beads will bind dsDNA >200bp). After binding DNA for 5 minutes, tubes are placed on a magnetic stand and supernatant is

discarded (or transferred to a different tube in case of upper cut-off size selection). Beads are then washed twice with 80% EtOH before DNA is eluted from the beads by adding the desired volume of water or 10mM Tris HCl pH 8.

Sequencing

All TCR cDNA libraries have been sequenced on the Nova-seq 6000 platform by Illumina using S4 2x150bp v1.5 kits with the following sequencing-cycle set-up: Read1: 150 cycles, Index1: 17 cycles, Read2: 150 cycles and Index2: 8 cycles.

Cost of CITR-seq experiments

In CITR-seq all molecular reactions are carried out in bulk for ~5.000-50.000 cells depending on the protocol step. This offers significant cost advantages, especially in contrast to plate-based single-cell protocols in which all molecular reactions are done separately for each cell. Enzymes needed for one experiment (using 500.000 input cells) in our hands cost about 350$ (ligase, reverse-transcriptase, polymerase, RNAse inhibitor etc.). The required barcoding oligos can be bought in high quantities and are then sufficient for many CITR-seq runs bringing down the oligo costs to less than 50$ per experiment. Collectively, the cost for library preparation in each experiment is therefore roughly 400$.

Analysis

CITR-seq sequencing data pre-processing

Demultiplexing of fastq-files was done using a custom script, allowing one nucleotide mismatch in the cellular barcode sequence (relative to the barcode whitelist). Afterwards, adapter sequences were trimmed from the sequencing reads using *cutadapt* [67]. We then used *UMItools* [68] to extract the 4bp in-line barcode sequence from each sequencing read. For each read the in-line barcode and the barcode sequence extracted from the corresponding index reads were combined. The combined barcode sequences were then added to the 5' end of read1. Afterwards, the full barcode information is present at the beginning of read 1 (16bp) followed by the UMI (10bp) and the 150bp sequencing

read. Read 2 contains just the 150bp sequencing read. This pre-processing of sequencing reads modifies the fastq-files to be easily integrated into the subsequent MiXCR-pipeline.

Species-specific V(D)J reference libraries

To construct individual V(D)J reference libraries for PWD/PhJ, CAST and SPRET we built on the strategy used in the *findAlleles* function implemented in the MiXCR [47] software. First, we used full-length TCR sequencing data of each species generated using the 10x Genomics Immune Profiling Kit (see below), to assemble gene-segment candidate-alleles: Raw sequencing fastq files were processed using *Cellranger VDJ* supplying the built-in mm10 based VDJ-reference (GRCm38-ensemble-7.0.0). In this pipeline fragmented reads are combined into full length contigs based on sequence overlap in reads and matching cellular barcodes. We used the generated "filtered_contig.fastq" output and passed it directly to the MiXCR alignment step ("*align*", --species mmu, --preset generic-amplicon --floating-left-alignment-boundary --floating-right-alignment-boundary C --rna) to generate binary vdjca-files. We then used *mixcr exportAlignments* (--dont-impute-germline-on-export -allNFeatures UTR5Begin FR3End) to extract gene-features so that SNPs in candidate-alleles are not modified to match the provided reference. For each candidate V(D)J-allele we then used the extremely unique combination of associated UMI and CDR3 sequences to distinguish low-frequency alleles from alleles generated by sequencing or PCR errors by requiring each allele to be identified with at least two unique CDR3/UMI combinations. The list of identified V,D and J segment alleles was then used to generate a MiXCR compatible reference libraries for each species using the *buildLibrary* function implemented in MiXCR. Since the underlying RNA-based input libraries are generated using template-switching rather than multiplex-PCR, they allow for the discovery of *de novo* V(D)J-segments since template-switch based cDNA libraries do not require previous knowledge of the entire set of gene-segments for amplification.

Full list of V(D)J genes/families analyzed in cross-species comparisons

All names of V(D)J genes/families correspond to the official IMGT nomenclature [57]. Pseudogenes as well as extremely low-expressed genes (< 200 transcripts across all ~5x10^6 T cells of all species) are excluded from the analysis. *Trbv24* (all species) and *Trbv31* (PWD) were excluded from the analysis due to failure of amplification during the multiplex PCR. The remaining list contains the following V(D)J genes/families:

1) **Trav-families:**
   Trav1, Trav2, Trav3, Trav4, Trav5, Trav6, Trav7, Trav8, Trav9, Trav10, Trav11, Trav12, Trav13, Trav14, Trav15, Trav16, Trav17, Trav18, Trav19, Trav21

2) **Trbv-genes:**
   Trbv1, Trbv2, Trbv3, Trbv4, Trbv5, Trbv12-1, Trbv12-2, Trbv13-1, Trbv13-2, Trbv13-3, Trbv14, Trbv15, Trbv16, Trbv17, Trbv19, Trbv20, Trbv21, Trbv23, Trbv26, Trbv29, Trbv30, Trbv31

3) **Traj-genes:**
   Traj2, Traj4, Traj5, Traj6, Traj7, Traj9, Traj11, Traj12, Traj13, Traj15, Traj16, Traj17, Traj18, Traj21, Traj22, Traj23, Traj24, Traj26, Traj27, Traj28, Traj30, Traj31, Traj32, Traj33, Traj34, Traj35, Traj37, Traj38, Traj39, Traj40, Traj42, Traj43, Traj44, Traj45, Traj47, Traj48, Traj49, Traj50, Traj52, Traj53, Traj54, Traj56, Traj57, Traj58

4) **Trbj-genes:**
   Trbj1-1, Trbj1-2, Trbj1-3, Trbj1-4, Trbj1-5, Trbj1-6, Trbj2-1, Trbj2-2, Trbj2-3, Trbj2-4, Trbj2-5, Trbj2-16, Trbj2-7

All gene names in the generated species-specific V(D)J reference files correspond to the closest relative (by sequence identity) in mm10 based MiXCR reference library.

Alignment of sequencing reads using MiXCR

Sequencing reads in pre-processed fastq-format were integrated into a custom MiXCR pipeline (MiXCR version 4.5.0) using the following steps:

4) *mixcr align*
   `--preset generic-ht-single-cell-amplicon-with-umi`
   `--library` Species Specific custom library (see above)
   `--tag-pattern ^(CELL:N(16))(UMI:N(10))(R1:*)\^(R2:*)`
   `--floating-left-alignment-boundary`
   `--floating-right-alignment-boundary C`
   `-OvParameters.geneFeatureToAlign=VRegionWithP`
   `-OminSumScore=100`

5) *mixcr refineTagsAndSort*

6) *mixcr assemble*
```
-- assemble-clonotypes-by CDR3
-- cell-level
```

We then used *mixcr exportClones* to extract the required information for all downstream analysis (e.g., cellular barcodes, transcript counts, V(D)J segments, CDR3 amino acid and nucleotide sequence etc.).

Construction of 10x Genomics Single Cell Immune Profiling sequencing libraries

We generated four sequencing libraries (from 10-week-old male mice, primary $CD8^+$ T cells of one of each: BL6, PWD, CAST, SPRET, see cell isolation described above) using the 10x Genomics Immune Profiling platform (Chromium Next GEM Single Cell 5' Kit v2) according to the manufacturer's instructions. T cells from each mouse were used in two separate reactions, each with 2.500 input cells (eight total reactions). V(D)J sequencing libraries were sequenced at 5.000 reads/cell. Raw sequencing data was pre-processed as described above and then aligned to species-specific V(D)J references using the outlined MiXCR pipeline.

Assignment of parental alleles in F1 hybrids

Pre-processed fastq-files of all F1 hybrid samples were aligned using MiXCR as described above. Importantly, the F1 hybrid samples were aligned to both parental V(D)J references and the alignment scores for V- and J-genes were extracted (*mixcr exportAlignments -vHitScore* and *-jHitScore*). We then compared the alignment scores for V- and J-genes from both alignments for each sequencing read. Each gene segment was then assigned to one parental species based on the higher alignment score in both alignments. Absence of SNPs in a gene-segment lead to identical alignment scores and therefore the respective reads were only assigned to a parental allele if the second gene segment in the same read was assigned to one parental allele. Reads in which both V- and J-segments had identical alignment scores in both alignments (e.g. no parental SNPs in both gene segments) as well as reads in which V- and J-parental assignment disagreed

were discarded from the analysis together with all other reads sharing the respective identical cellular barcode.

Comparison of CITR-seq data with publicly available datasets from Parse Bioscience and 10x Genomics

Absolute counts of paired αβ-TCRs shown in **Fig. 2D** were taken from the following datasets:

1) **Parse Bioscience [44]:**
   TCR Sequencing of 1 Million Primary Human T Cells in a Single Experiment (primary human Pan T cells, sequencing depth: 5000 reads/cell).

2) **10x Genomics Single Cell Immune Profiling [45]:**
   CD8+ T cells of Healthy Donor 2 (v1, 150x91), Single Cell Immune Profiling Dataset by Cell Ranger v3.0.2, 10x Genomics, (2019, May, 9)

The UMI/cell recovery rates in CITR-seq were compared to the UMI/cell recovery rates in two publicly available datasets provided by Parse Bioscience:

1) **Parse Bioscience [44]:**
   TCR Sequencing of 1 Million Primary Human T Cells in a Single Experiment (primary human Pan T cells, sequencing depth: 5000 reads/cell)

2) **Parse Bioscience [46]:**
   Performance of Evercode TCR in Activated Human T cells
   (Pan T cells after 72h activation using CD3/CD28 beads + IL-2 supplementation, sequencing depth: 5000 reads/cell)

Datasets are available from their website (https://www.parsebiosciences.com) and the specific UMI/cell rates were extracted from the "TCR:Barcode Report (TSV)" tables (column: "transcript_count")

PCA of VJ-pairing and central CDR3 4mer abundance

We conducted Principal Component Analysis (PCA) on two different datasets.

1) **V-J pairing**

The first PCA was done to compare total counts of observed V-J pairs across all sample down-sampled to a common cell count of 5.000 T cells (each with one associated TCRα and TCRβ chain). The count-tables were analyzed using DESeq2's[69] *varianceStabilizingTransformation* (*vst*, blind=FALSE, nsub=300) and PCA was conducted on the top 300 most variable V-J pairs (*plotPCA* , ntop=300). Using this parameters PC1-4 explain approximately 67% of variance in V-J usage across samples.

2) **CDR3 4mers**

The second PCA analysis was done on a set of amino acid 4mers (or 4mer pairs) extracted from the central region of CDR3 amino acid motifs (the three most 3' and 5' amino acids were trimmed from the motif). Initially three randomly chosen 4mers were extracted from each trimmed CDR3 motif. For paired CDR3αβ motifs we extracted 3 random 4mer pairs of the respective CDR3α and CDR3β motifs associated with a cellular barcode. Subsequently, we generated count-matrices with the total counts of each 4mer across the 32 individuals. We filtered the matrices to only contain 4mer/pairs that were observed at least once across three individuals of a specific genotype (final 4mer counts: 39.843 CDR3α, 56.538 CDR3β and 1.201.646 CDR3αβ 4mers). The filtered count matrices were then analyzed following the standard DESeq2 [69] workflow for un-normalized count-matrix inputs. Afterwards, PCA was done on the top 5000 most variable 4mers using the *plotPCA* function.

Diversity and overlap indices used for repertoire comparison in CITR-seq data

Many of the commonly used indices from the analysis of TCR repertoires within and across different samples, were originally developed to quantify the diversity of species in an ecosystem. For this reason, they are often classified as either *alpha-diversity* indices that measure species richness and/or evenness within a particular population or alternatively as *beta-diversity* indices, which evaluate differences or overlaps between different populations. Similar to species diversity studies in ecology, TCR repertoire diversity estimates suffer from inherent incompleteness of the sampled diversity, a problem first described as the unseen species problem [70]. The use of diversity indices for adaptive immune receptor analysis is reviewed here [71]. The following indices were used in this study:

1) **Shannon diversity index** [72] (for proportions)

   The Shannon diversity index considers both, species richness and species evenness to evaluate the entropy within a distribution of species:

$$H = -\sum_{i=1}^{k} p_i \log(p_i)$$

   With p_i = the proportion (frequency) in the group k (e.g. gene segments). The index can be normalized by dividing it by the maximum diversity. Which then is the normalized Shannon diversity index (nSDI) used in this study:

$$E_H = \frac{H}{\log(k)}$$

   In the context of gene segment usage in nSDI of 1 would indicate that all gene segments are used at identical frequencies in a TCR repertoire.

2) **Jaccard index**

The Jaccard index was developed by Paul Jaccard in 1901 and is commonly used to calculate the overlap (of CDR3 motifs) between two samples of TCRs:

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

The Jaccard index calculates the intersection size divided by union size of two samples (A and B).

Classification of relative V(D)J gene usage in parental lines and their F1 hybrids

In classical F1 hybrid experiments, genes are often categorized into additive, dominant, over- and under-dominant, based on their expression in F1 hybrids relative to the parental individuals [73]. When adopted to V(D)J-gene usage in F1 hybrids and their parental lines it is important to note, that the frequency of a particular gene does not only depend on differences in gene expression regulation but is also influenced by biases during V(D)J recombination and thymic selection. We see that thymic selection introduces significant changes to V(D)J gene usage and therefore amplifies gene usage differences across species relative to the differences emerging from differential gene regulation alone. This effect is especially strong in F1 hybrids were almost all V(D)J genes show significantly different frequencies relative to the parental species (**Fig S5A**). Instead of using a *p*-value based classification, we therefore decided to rather compare the relative frequencies of V(D)J gene usage across F1 hybrids and the parental species. Accordingly, V(D)J gene frequencies are classified using the following criteria:

1) **Conserved:** Gene frequency in the F1 hybrid is within 1% of the frequency in both parents

2) **Dominant:** Gene frequency in the F1 hybrid is within 1% of the frequency in one parent and more than 1% larger or smaller than the frequency in the other parent.

3) **Additive:** Gene frequency in the F1 hybrid is more than 1% smaller than the frequency in one parent and more than 1% larger than the frequency in the other parent.

4) **Over-dominant:** Gene frequency in the F1 hybrid is more than 1% larger than the frequency in bother parents.

5) **Under-dominant:** Gene frequency in the F1 hybrid is more than 1% smaller than the frequency in bother parents.

Acknowledgements

We thank all past and present members of the Chan and Jones laboratory for input into experimental design, helpful discussion and improving the manuscript. We especially thank Felicity Jones for her scientific input throughout the entire study. We thank Sinja Mattes and the remaining team of animal caretakers at the Friedrich-Miescher Laboratory led by Cemal Yilmaz. We thank Aurora Panzera and Christian Feldhaus from the Bio Optics core facility of the Max Planck Institute for Biology for their support with all FACS experiments. We thank Insa Hirschberg for supporting tissue culture experiments. We also thank the Genome Center of the Max Planck Institute for Biology for providing support with CITR-seq library sequencing. M.P. and D.S. are supported by an International Max Planck Research School fellowship. M.K. and Y.F.C are supported by the European Research Council Starting Grant 639096 “HybridMix” and Proof-of-Concept Grant 101069216 “Haplotagging”. The research done in this study is supported by the Max Planck Society.

Author Contributions

M.P. and Y.F.C. designed the experiments. M.P. and V.S. developed the barcoding framework for CITR-seq. M.P. developed the rest of the protocol and performed all experiments. M.P. performed the computational analysis advised by Y.F.C. M.P. wrote the manuscript. V.S., D.S., M.K. and Y.F.C. provided support for the experiments and the computational analysis. All authors reviewed the manuscript. Y.F.C. direct the study.

Declaration of Interest

The authors declare no competing interests.

Literature

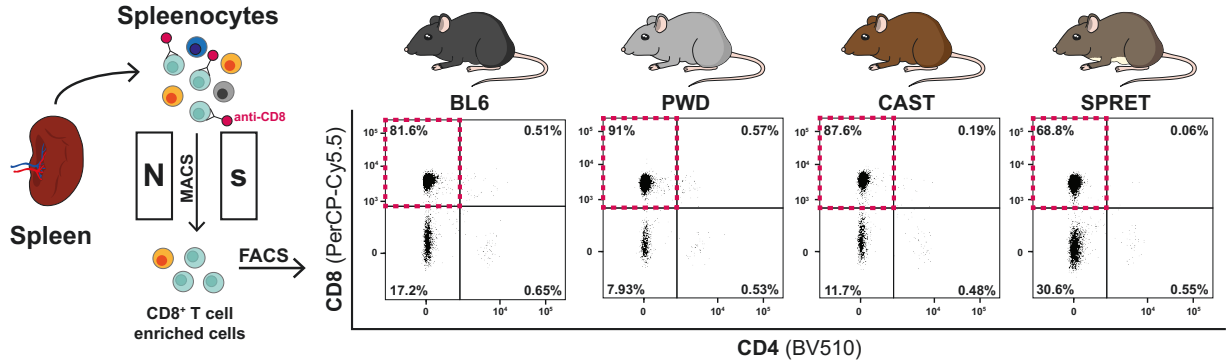
1. von Boehmer, H. and P. Kisielow, *Self-nonself discrimination by T cells.* Science, 1990. **248**(4961): p. 1369-73.
2. Raskov, H., et al., *Cytotoxic CD8(+) T cells in cancer and cancer immunotherapy.* Br J Cancer, 2021. **124**(2): p. 359-367.
3. Luckheeram, R.V., et al., *CD4(+)T cells: differentiation and functions.* Clin Dev Immunol, 2012. **2012**: p. 925135.
4. Davis, M.M. and P.J. Bjorkman, *T-cell antigen receptor genes and T-cell recognition.* Nature, 1988. **334**(6181): p. 395-402.
5. Nadel, B. 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, 1997. **17**(7): p. 3768-78.
6. Glanville, J., et al., *Identifying specificity groups in the T cell receptor repertoire.* Nature, 2017. **547**(7661): p. 94-98.
7. Garcia, K.C., et al., *Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen.* Science, 1998. **279**(5354): p. 1166-72.
8. Wu, L.C., et al., *Two-step binding mechanism for T-cell receptor recognition of peptide MHC.* Nature, 2002. **418**(6897): p. 552-6.
9. Dutta, A., B. Zhao, and P.E. Love, *New insights into TCR β-selection.* Trends in Immunology, 2021. **42**(8): p. 735-750.
10. Klein, L., et al., *Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see).* Nat Rev Immunol, 2014. **14**(6): p. 377-91.
11. Sethna, Z., et al., *Population variability in the generation and selection of T-cell repertoires.* PLoS Comput Biol, 2020. **16**(12): p. e1008394.
12. Krishna, C., et al., *Genetic and environmental determinants of human TCR repertoire diversity.* Immun Ageing, 2020. **17**: p. 26.
13. Dewitt, W.S., et al., *Human T cell receptor occurrence patterns encode immune history, genetic background, an receptor specificity.* Elife, 2018. **7**.
14. Tanno, H., et al., *Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins.* Proceedings of the National Academy of Sciences of the United States of America, 2020. **117**(1): p. 532-540.
15. Glanville, J., et al., *Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation.* Proc Natl Acad Sci U S A, 2011. **108**(50): p. 20066-71.
16. Zvyagin, I.V., et al., *Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing.* Proc Natl Acad Sci U S A, 2014. **111**(16): p. 5980-5.
17. Garcia, K.C., *Reconciling views on T cell receptor germline bias for MHC.* Trends Immunol, 2012. **33**(9): p. 429-36.

18. Rangarajan, S. and R.A. Mariuzza, *T cell receptor bias for MHC: co-evolution or co-receptors?* Cell Mol Life Sci, 2014. **71**(16): p. 3059-68.
19. Van Laethem, F., et al., *Deletion of CD4 and CD8 coreceptors permits generation of alphabetaT cells that recognize antigens independently of the MHC.* Immunity, 2007. **27**(5): p. 735-50.
20. La Gruta, N.L., et al., *Understanding the drivers of MHC restriction of T cell receptors.* Nat Rev Immunol, 2018. **18**(7): p. 467-478.
21. Feng, D., et al., *Structural evidence for a germline-encoded T cell receptor-major histocompatibility complex interaction 'codon'.* Nat Immunol, 2007. **8**(9): p. 975-83.
22. Pierini, F. and T.L. Lenz, *Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection.* Molecular Biology and Evolution, 2018. **35**(9): p. 2145-2158.
23. Penn, D.J., K. Damjanovich, and W.K. Potts, *MHC heterozygosity confers a selective advantage against multiple-strain infections.* Proc Natl Acad Sci U S A, 2002. **99**(17): p. 11260-4.
24. Migalska, M., A. Sebastian, and J. Radwan, *Major histocompatibility complex class I diversity limits the repertoire of T cell receptors.* Proc Natl Acad Sci U S A, 2019. **116**(11): p. 5021-5026.
25. Woelfing, B., et al., *Does intra-individual major histocompatibility complex diversity keep a golden mean?* Philos Trans R Soc Lond B Biol Sci, 2009. **364**(1513): p. 117-28.
26. Janeway, C. and C. Janeway, *Immunobiology : the immune system in health and disease*. 5th ed. 2001, New York: Garland Pub. xviii, 732 pages : illustrations.
27. Zarnitsyna, V.I., et al., *Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire.* Front Immunol, 2013. **4**: p. 485.
28. Thierry Mora, A.W., *Quantifying lymphocyte receptor diversity.* Systems Immunology, (ed. J. Das, C. Jayaprakash), 2019.
29. Arstila, T.P., et al., *A direct estimate of the human alphabeta T cell receptor diversity.* Science, 1999. **286**(5441): p. 958-61.
30. Casrouge, A., et al., *Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes.* J Immunol, 2000. **164**(11): p. 5782-7.
31. Wiegel, F.W. and A.S. Perelson, *Some scaling principles for the immune system.* Immunology and Cell Biology, 2004. **82**(2): p. 127-131.
32. Langman, R.E. and M. Cohn, *The E-T (elephant-tadpole) paradox necessitates the concept of a unit of B-cell function: the protection.* Mol Immunol, 1987. **24**(7): p. 675-97.
33. Freeman, J.D., et al., *Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing.* Genome Res, 2009. **19**(10): p. 1817-24.
34. Robins, H.S., et al., *Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells.* Blood, 2009. **114**(19): p. 4099-107.
35. Wang, C., et al., *High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets.* Proc Natl Acad Sci U S A, 2010. **107**(4): p. 1518-23.
36. Han, A., et al., *Corrigendum: Linking T-cell receptor sequence to functional phenotype at the single-cell level.* Nat Biotechnol, 2015. **33**(2): p. 210.
37. Neal, J.T., et al., *Organoid Modeling of the Tumor Immune Microenvironment.* Cell, 2018. **175**(7): p. 1972-1988 e16.
38. Tu, A.A., et al., *TCR sequencing paired with massively parallel 3' RNA-seq reveals clonotypic T cell signatures.* Nat Immunol, 2019. **20**(12): p. 1692-1699.
39. Rosenberg, A.B., et al., *Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding.* Science, 2018. **360**(6385): p. 176-182.
40. Goios, A., et al., *mtDNA phylogeny and evolution of laboratory mouse strains.* Genome Res, 2007. **17**(3): p. 293-8.
41. Morgan, A.P., et al., *Population structure and inbreeding in wild house mice (Mus musculus) at different geographic scales.* Heredity (Edinb), 2022. **129**(3): p. 183-194.
42. Yang, H., et al., *Subspecific origin and haplotype diversity in the laboratory mouse.* Nat Genet, 2011. **43**(7): p. 648-55.
43. Howie, B., et al., *High-throughput pairing of T cell receptor alpha and beta sequences.* Sci Transl Med, 2015. **7**(301): p. 301ra131.
44. Bioscience, P., *TCR Sequencing of 1 Million Primary Human T Cells in a Single Experiment*, P. Bioscience, Editor.

45. Genomics, x., *CD8+ T cells of Healthy Donor 2, (v1, 150x91), Single Cell Immune Profiling Dataset*, x. Genomics, Editor. 2019.
46. Bioscience, P., *Performance of Evercode TCR in Activated Human T Cells*.
47. Bolotin, D.A., et al., *MiXCR: software for comprehensive adaptive immunity profiling.* Nat Methods, 2015. **12**(5): p. 380-1.
48. Bosc, N. and M.P. Lefranc, *The mouse (Mus musculus) T cell receptor alpha (TRA) and delta (TRD) variable genes.* Dev Comp Immunol, 2003. **27**(6-7): p. 465-97.
49. Hou, X.L., et al., *High Throughput Sequencing of T Cell Antigen Receptors Reveals a Conserved TCR Repertoire.* Medicine, 2016. **95**(10).
50. Kitaura, K., et al., *A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) alpha and beta repertoires and identifying potential new invariant TCR alpha chains.* BMC Immunol, 2016. **17**(1): p. 38.
51. Thompson, S.D., J. Pelkonen, and J.L. Hurwitz, *First T cell receptor alpha gene rearrangements during T cell ontogeny skew to the 5' region of the J alpha locus.* J Immunol, 1990. **145**(7): p. 2347-52.
52. Mahowald, G.K., et al., *Out-of-Frame T Cell Receptor Beta Transcripts Are Eliminated by Multiple Pathways.* Plos One, 2011. **6**(7).
53. Weischenfeldt, J., et al., *NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements.* Genes & Development, 2008. **22**(10): p. 1381-1396.
54. Murugan, A., et al., *Statistical inference of the generation probability of T-cell receptors from sequence repertoires.* Proc Natl Acad Sci U S A, 2012. **109**(40): p. 16161-6.
55. Sharon, E., et al., *Genetic variation in MHC proteins is associated with T cell receptor expression biases.* Nature Genetics, 2016. **48**(9): p. 995-+.
56. Logunova, N.N., et al., *MHC-II alleles shape the CDR3 repertoires of conventional and regulatory naive CD4 T cells.* Proceedings of the National Academy of Sciences of the United States of America, 2020. **117**(24): p. 13659-13669.
57. Lefranc, M.P., et al., *IMGT®, the international ImMunoGeneTics information system®.* Nucleic Acids Research, 2009. **37**: p. D1006-D1012.
58. Pai, J.A. and A.T. Satpathy, *High-throughput and single-cell T cell receptor sequencing technologies.* Nat Methods, 2021. **18**(8): p. 881-892.
59. Huang, C.Y. and O. Kanagawa, *Ordered and coordinated rearrangement of the T cell receptor α locus:: role of secondary rearrangement in thymic selection.* Faseb Journal, 2001. **15**(5): p. A1023-A1023.
60. Lawlor, D.A., et al., *Evolution of Class-I Mhc Genes and Proteins - from Natural-Selection to Thymic Selection.* Annual Review of Immunology, 1990. **8**: p. 23-63.
61. Vidovic, D. and P. Matzinger, *Unresponsiveness to a Foreign Antigen Can Be Caused by Self-Tolerance.* Nature, 1988. **336**(6196): p. 222-225.
62. Arrieta-Bolanos, E., D.I. Hernandez-Zaragoza, and R. Barquera, *An HLA map of the world: A comparison of HLA frequencies in 200 worldwide populations reveals diverse patterns for class I and class II.* Front Genet, 2023. **14**: p. 866407.
63. Argaet, V.P., et al., *Dominant Selection of an Invariant T-Cell Antigen Receptor in Response to Persistent Infection by Epstein-Barr-Virus.* Journal of Experimental Medicine, 1994. **180**(6): p. 2335-2340.
64. Day, E.K., et al., *Rapid CD8T cell repertoire focusing and selection of high-affinity clones into memory following primary infection with a persistent human virus:: Human Cytomegalovirus.* Journal of Immunology, 2007. **179**(5): p. 3203-3213.
65. Trautmann, L., et al., *Selection of T cell clones expressing high-affinity public TCRs within human cytomegalovirus-specific CD8 T cell responses.* Journal of Immunology, 2005. **175**(9): p. 6123-6132.
66. Lewis, M.D., et al., *A reproducible method for the expansion of mouse CD8+ T lymphocytes.* J Immunol Methods, 2015. **417**: p. 134-138.
67. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal, 2011. **17(1)**: p. pp. 10-12.
68. Smith, T., A. Heger, and I. Sudbery, *UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy.* Genome Res, 2017. **27**(3): p. 491-499.

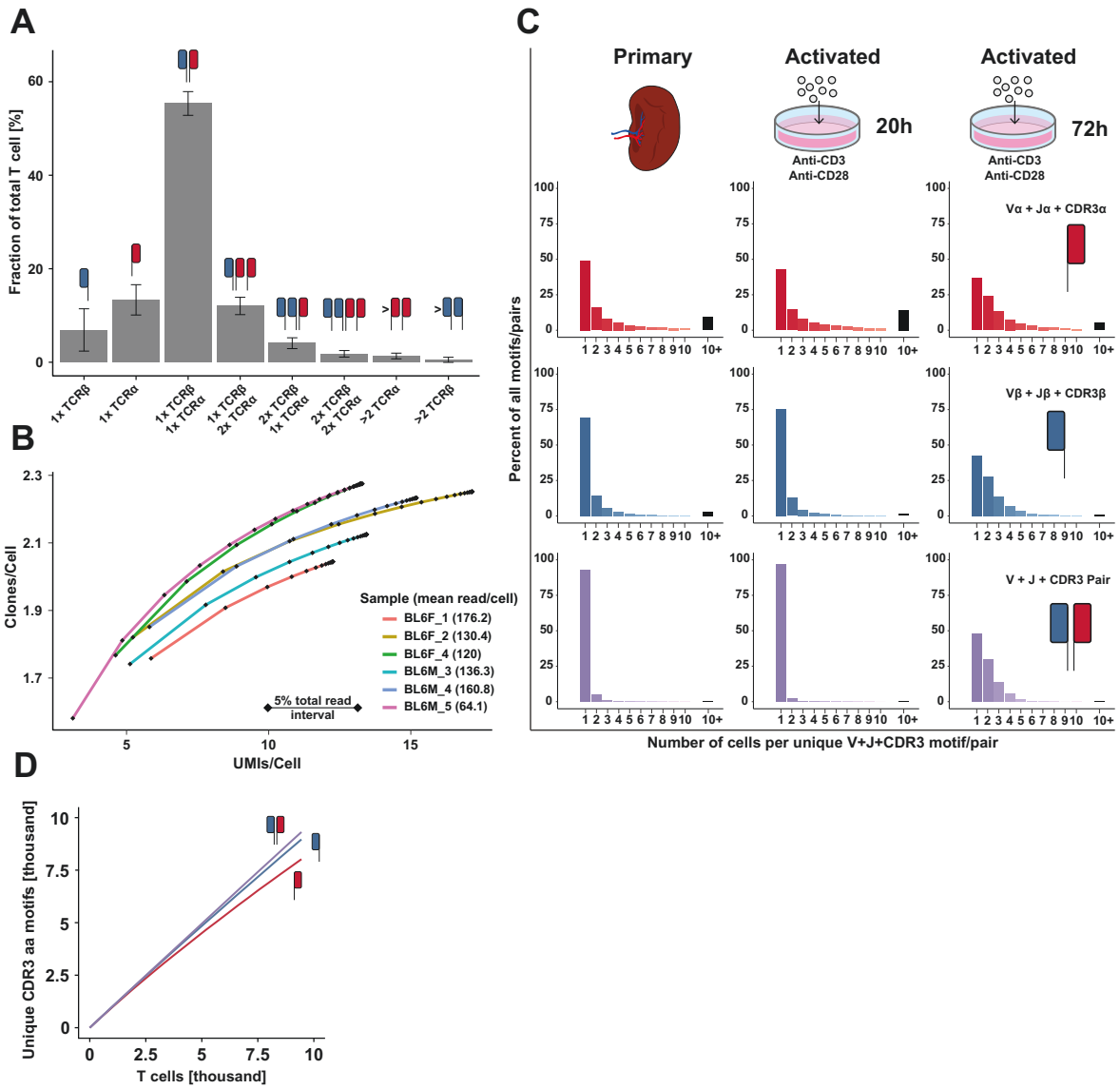
69. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12).
70. Fisher, R.A., A.S. Corbet, and C.B. Williams, *The relation between the number of species and the number of individuals in a random sample of an animal population.* Journal of Animal Ecology, 1943. **12**: p. 42-58.
71. Chiffelle, J., et al., *T-cell repertoire analysis and metrics of diversity and clonality.* Curr Opin Biotechnol, 2020. **65**: p. 284-295.
72. Shannon, C.E., *A Mathematical Theory of Communication.* Bell System Technical Journal, 1948. **27**(3): p. 379-423.
73. Metzger, B.P.H., P.J. Wittkopp, and J.D. Coolon, *Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Species.* Genome Biology and Evolution, 2017. **9**(4): p. 843-854.

Supplement



Supplementary Figure 1: CD8⁺ T cells isolation strategy for all CITR-seq samples

(A) Spleenocyte cell-suspensions were pre-enriched for CD8⁺ T cells by magnetic extraction of anti-CD8 labeled cells (magnetic-activated cell sorting, MACS using Dynabeads™ FlowComp™ Mouse CD8 Kit). Afterwards pre-enriched cell-suspension was further purified using fluorescence activated cell sorting (FACS). Percentages in each quadrant of the FACS plots represent the mean frequencies of the respective cell population in the pre-enriched cell-suspension. CD8⁺ T cells in the top left quadrant (red box) were sorted and used for CITR-seq experiments.



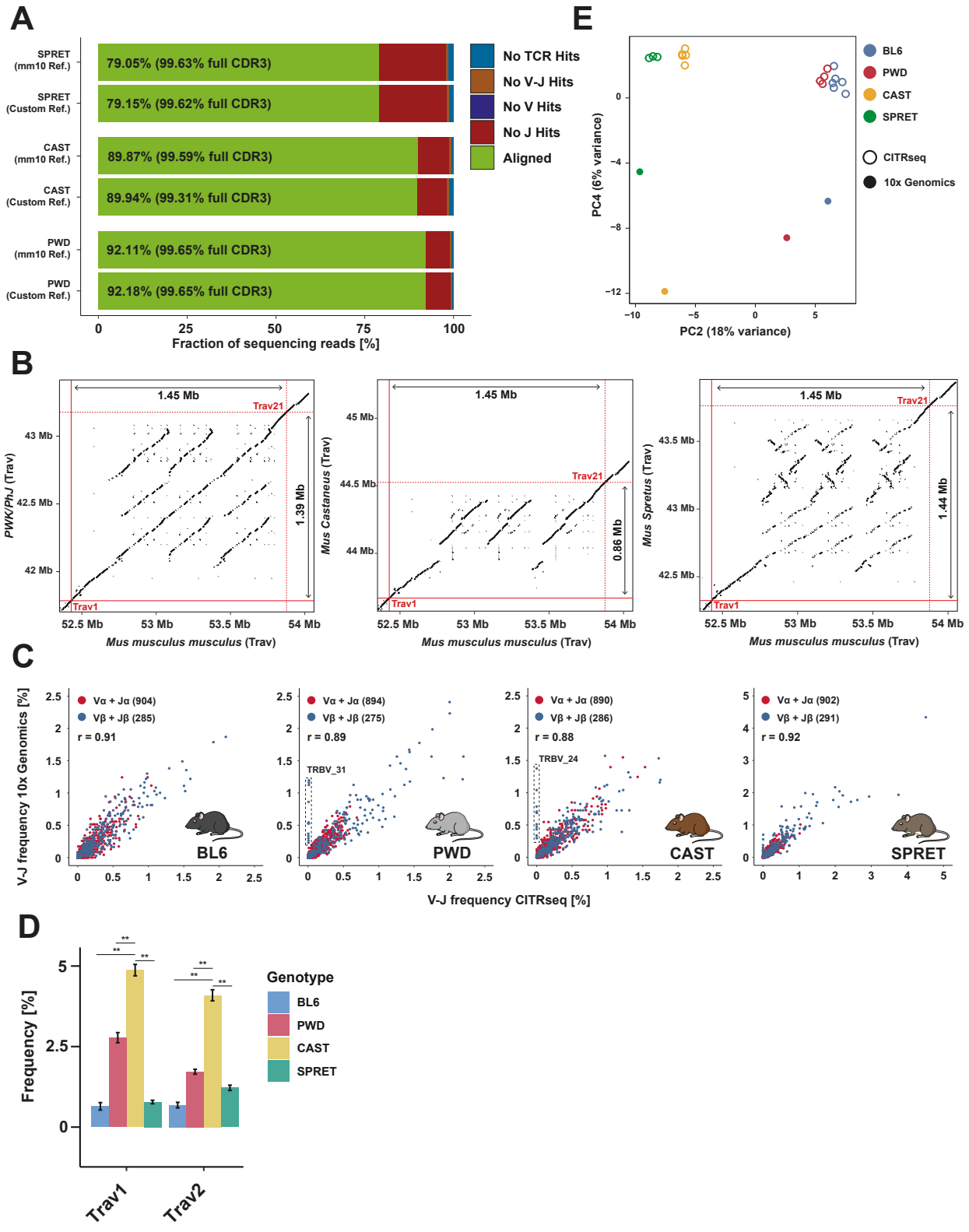
Supplementary Figure 2: Additional analysis for CITR-seq validation

(A) Mean fraction of T cells assigned to different numbers of distinct TCRα and TCRβ chains (error bars represent the standard deviation across all 32 CITR-seq samples). Most T cells (~55%) are associated with a single TCRα and a single TCRβ chain. Few T cells are associated with more than two TCRα (~1.4%) or TCRβ (~0.5%) chains, likely representing cell doublets or barcode collisions.

(B) Saturation curve showing UMI/cell and clone/cell counts relative to the fraction of total sequencing reads. Diamonds represent the respective UMI/cell and clone/cell counts at intervals of 5% of sequencing reads (5% - 100% of reads) for six representative CITR-seq samples (all BL6 samples). The mean reads per cell are shown for the representative samples.

(C) Clone size distributions (number of cells observed with a unique V+J+CDR3 TCR) in samples from primary T cells (left), 20h activated T cells (middle) and 72h activated T cells (right). The respective clone size distributions are shown for Vα+Jα+CDR3α TCRs (top), Vβ+Jβ+CDR3β TCRs (middle) or V+J+CDR3 paired αβ-TCRs (bottom). In contrast to primary and 20h activated T cells, 72h activated T cells show an increased clone size distribution caused by the onset of clonal expansion by prolonged T cell activation.

(D) Total number of unique CDR3α, CDR3β or paired CDR3αβ amino acid motifs relative to the number of T cells across all 4 samples generated using the 10x Genomics Single Cell Immune Profiling Kit.


A
SPRET (mm10 Ref.)
SPRET (Custom Ref.)
CAST (mm10 Ref.)
CAST (Custom Ref.)
PWD (mm10 Ref.)
PWD (Custom Ref.)
79.05% (99.63% full CDR3)
79.15% (99.62% full CDR3)
89.87% (99.59% full CDR3)
89.94% (99.31% full CDR3)
92.11% (99.65% full CDR3)
92.18% (99.65% full CDR3)
No TCR Hits
No V-J Hits
No V Hits
No J Hits
Aligned
Fraction of sequencing reads [%]
E
PC4 (6% variance)
PC2 (18% variance)
BL6
PWD
CAST
SPRET
CITRseq
10x Genomics
B
1.45 Mb
Trav21
Trav1
1.39 Mb
0.86 Mb
1.44 Mb
PWK/PhJ (Trav)
Mus Castaneus (Trav)
Mus Spretus (Trav)
Mus musculus musculus (Trav)
C
Vα + Jα (904)
Vβ + Jβ (285)
r = 0.91
Vα + Jα (894)
Vβ + Jβ (275)
r = 0.89
TRBV_31
Vα + Jα (890)
Vβ + Jβ (286)
r = 0.88
TRBV_24
Vα + Jα (902)
Vβ + Jβ (291)
r = 0.92
BL6
PWD
CAST
SPRET
V-J frequency 10x Genomics [%]
V-J frequency CITRseq [%]
D
Frequency [%]
Genotype
BL6
PWD
CAST
SPRET
Trav1
Trav2

Supplementary Figure 3: Different structure of TCRα loci across inbred species and comparison of observed V-J usage frequencies in different methods

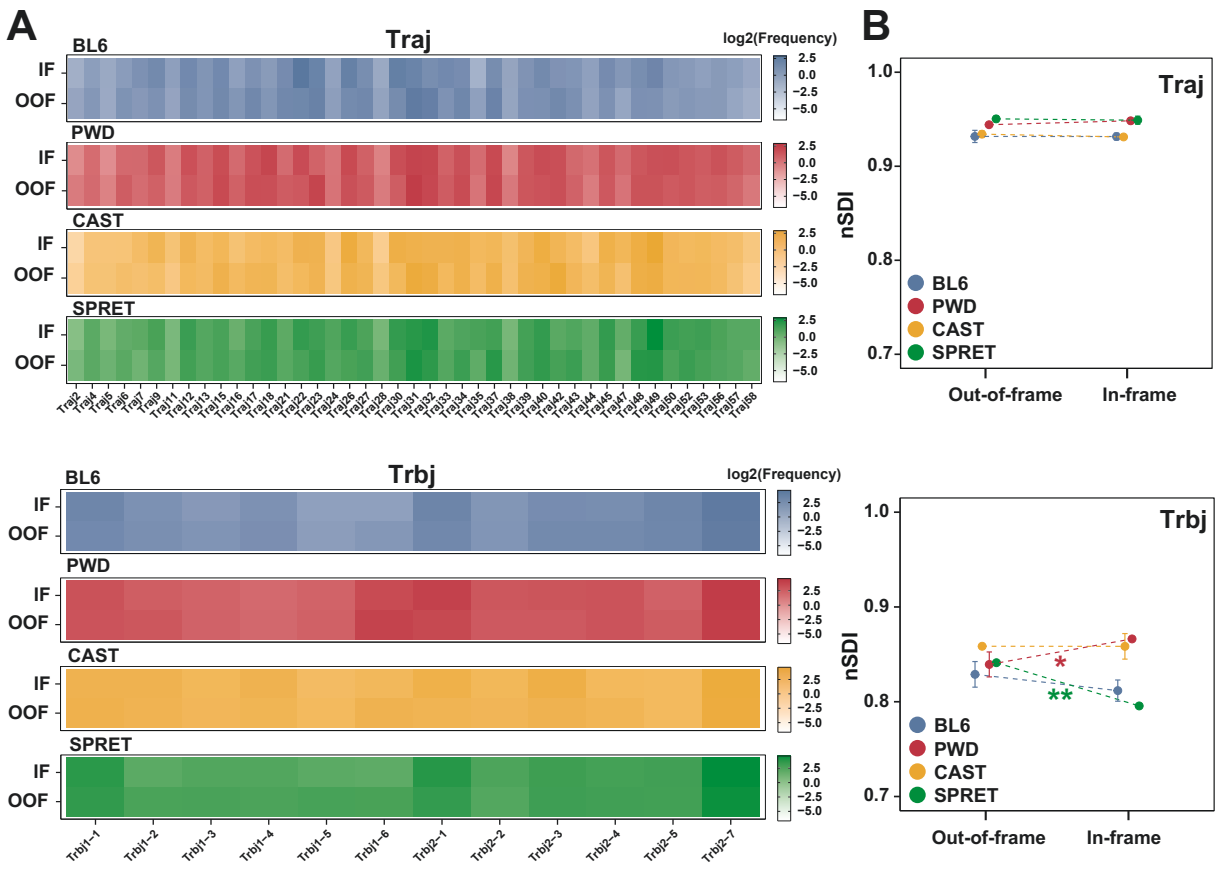
(A) Fraction of sequencing reads that were successfully aligned to V(D)J genes using different reference libraries (green bars; full CDR3 coverage in brackets). Each stacked bar shows the mapping percentage for a representative sample from SPRET, CAST and PWD when aligned to the in-build mm10 based MiXCR V(D)J reference (top) and the species-specific custom V(D)J reference (see methods). All other colors in the stacked bar represent the reason for the failure of alignment. In all cases, the total fraction of successfully aligned reads is higher when using the species-specific custom library.

(B) Dot plots of local alignment of genomic sequence from the GRCm38/mm10 TCR Vα locus to the PWK/PhJ (closest available genomic sequence to PWD, left), CAST (middle) and SPRET (right) genomic sequence of the TCR Vα locus. Intersections of the red lines indicate the location of the most distal (*Trav1*) and proximal (*Trav21*, dashed line) Vα genes. The genomic distances between these two Vα genes are shown. The central region of the Vα cluster is triplicated in BL6, PWK and SPRET relative to CAST.

(C) Dot plots showing the mean frequency of single-chain V-J pairing in TCRα (red) and TCRβ (blue) chains observed in samples generated with -seq and 10x Genomics Single Cell Immune Profiling. The respective frequencies are shown for BL6, PWD, CAST and SPRET samples. Pearson-correlation and the total number of detected Vα-Jα and Vβ-Jβ are shown. Boxes highlight Vβ genes that are almost exclusively observed in 10x Genomics samples indicating failure of amplification for these Vβ genes by the multiple PCR primer pool used in CITR-seq. The respective Vβ genes were excluded from the analysis.

(D) Usage frequencies of distal (5') Vα genes (*Trav1*, *Trav2)* with one-to-one orthology across all four species. CAST mice have significantly higher frequencies of both genes compared to all other species (chi-squared test, ** *P-value* < 0.01).

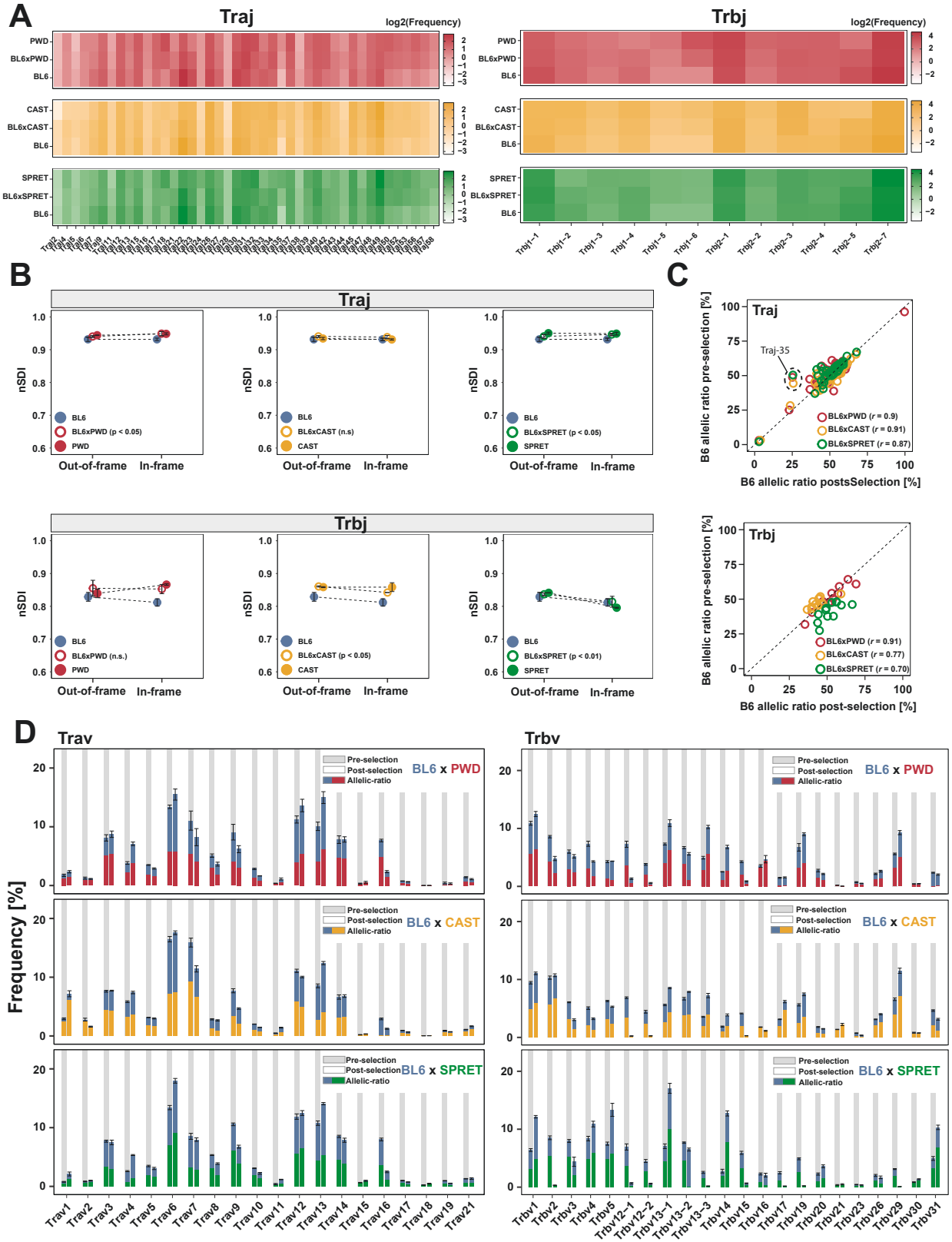
(E) PCA of combined Vα-Jα and Vβ-Jβ gene segment usage frequencies across all species as observed in samples generated with CITR-seq (empty circle) or 10x Genomics Single Cell Immune Profiling (filled circles). PC2 and PC4 are shown. PC4 contains 6% of the total variance across samples and separates samples by the respective methods used to generate the data.



Supplementary Figure 4: Comparison of Jα and Jβ gene usage in in-frame and out-of-frame TCRs

(A) Jα family (top) and Jβ gene (bottom) usage frequency (log_2) heatmaps. Heatmaps show the mean intra-species J-usage in in-frame (IF) and out-of-frame (OOF) TCRs across all T cells.

(B) Mean intra-species entropy in Jα-usage (top) and Jβ-usage (bottom) distributions calculated using the normalized Shannon diversity index (nSDI) for OOF and IF TCRs (error bars indicate the standard deviation in species replicates, significance calculated using paired t-test, * *P-value* < 0.05, ** *P-value* < 0.01).


A
Traj
log2(Frequency)
PWD
BL6xPWD
BL6
CAST
BL6xCAST
SPRET
BL6xSPRET
Trbj
B
nSDI
Out-of-frame
In-frame
BL6
BL6xPWD (p < 0.05)
PWD
BL6xCAST (n.s)
CAST
BL6xSPRET (p < 0.05)
SPRET
BL6xPWD (n.s.)
BL6xCAST (p < 0.05)
BL6xSPRET (p < 0.01)
C
B6 allelic ratio pre-selection [%]
B6 allelic ratio postsSelection [%]
Traj-35
BL6xPWD (r = 0.9)
BL6xCAST (r = 0.91)
BL6xSPRET (r = 0.87)
B6 allelic ratio post-selection [%]
BL6xPWD (r = 0.91)
BL6xCAST (r = 0.77)
BL6xSPRET (r = 0.70)
D
Trav
Trbv
Frequency [%]
Pre-selection
Post-selection
Allelic-ratio
BL6 x PWD
BL6 x CAST
BL6 x SPRET

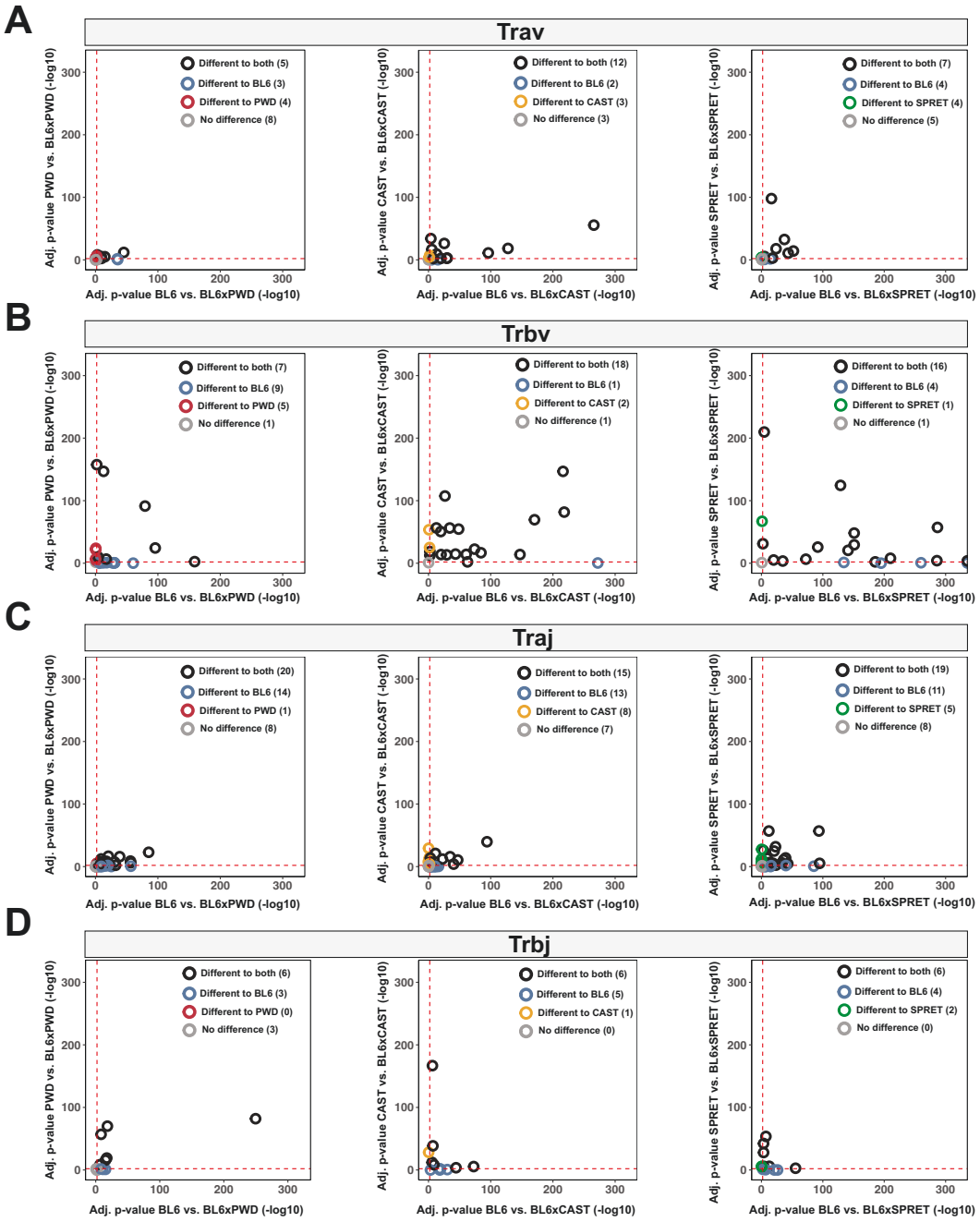
Supplementary Figure 5: Usage frequencies of Jα and Jβ genes in F1 hybrids and the impact of thymic selection on their abundance

(A) Jα gene (left) and Jβ gene (right) usage frequency (log_2) heatmaps of in-frame TCRs in F1 hybrids and their respective parental species.

(B) Comparison of entropy of J-usage distribution in F1 hybrids and the respective parental species calculated using the normalized Shannon diversity index (nSDI) for OOF (left) and IF (right) TCRs (error bars indicate the standard deviation in species replicates, significance tested for F1 hybrid IF vs OOF contrast using paired t-tests).

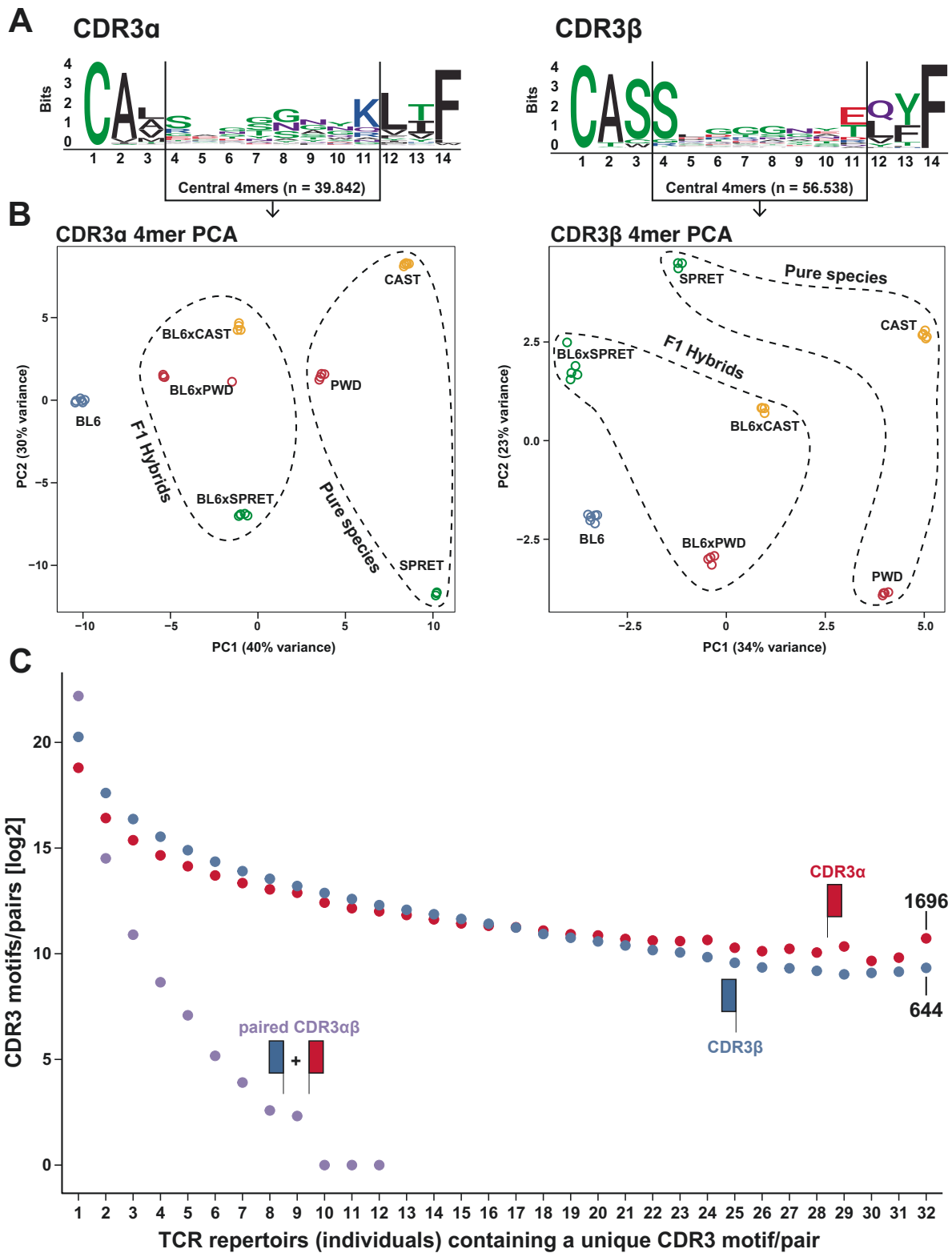
(C) Analysis of biased J gene allele usage in F1 hybrids. Plots show the percentage of BL6 Jα gene alleles and Jβ gene alleles in post- (x-axis) and pre-selection (y-axis) TCRs. Each circle represents a Jα-gene (top) or Jβ-gene (bottom). Pearson-correlation was calculated for post- and pre-selection J gene usage. Genes with substantial changes in allelic ratios in pre- and post-selection repertoires are highlighted (*Traj35*).

(D) Detailed representation of the mean Vα-family and Vβ-gene usage frequencies in F1 hybrids in pre-selection (grey background) and post-selection (white background) TCRs. Stacked bars show the allelic ratio in the respective V-gene/family (error bars indicate the standard deviation in species replicates).



Supplementary Figure 6: *P-values* (-log$_{10}$) of V/J gene usage frequency changes in F1 hybrids relative to both parents

Adjusted *P-values* for V and J gene frequency changes in F1 hybrids relative to their parents. Plots show *P-values* for BL6xPWD (left), BL6xCAST (middle) and BL6xSPRET (right) for Vα-families **(A)**, Vβ-genes **(B)**, Jα-genes **(C)** and Jβ-genes **(D)**. *P-values* have been calculated for differences in absolute count of TCRs with the respective V/J genes using Wald-test. Dashed red lines show that *P-value* cut-off of $P < 0.01$. Genes with significant ($P < 0.01$) changes relative to both parents (empty black circles), to the BL6 parent (empty blue circles), the respective other parent (PWD, CAST, SPRET, empty red, yellow, green circles) as well as genes with no differences to both parents (empty grey circles) are shown with the respective total counts of genes in each category.


A
CDR3α
CDR3β
Bits
Central 4mers (n = 39.842)
Central 4mers (n = 56.538)
B
CDR3α 4mer PCA
CDR3β 4mer PCA
BL6
BL6xCAST
BL6xPWD
BL6xSPRET
CAST
PWD
SPRET
F1 Hybrids
Pure species
PC1 (40% variance)
PC2 (30% variance)
PC1 (34% variance)
PC2 (23% variance)
C
CDR3 motifs/pairs [log2]
TCR repertoirs (individuals) containing a unique CDR3 motif/pair
paired CDR3αβ
CDR3α
CDR3β
1696
644

Supplementary Figure 7: Comparison of single-chain CDR3α and CDR3β amino acid motifs in all 32 CITR-seq samples

(A) Positional diversity of amino acids in 10,000 randomly chosen CDR3α (left) and CDR3β (right) single-chain motifs from all 32 CITRs-eq samples. Black lines mark the central region of CDR3 motifs. Three 4mers were randomly selected from this central region of all CDR3 sequences across the 32 individuals. The filtered count-matrix of 4mers (see methods) contained 39,842 unique CDR3α and 56,538 unique CDR3β 4mers.

(B) PCA analysis done using the 4mer count-matrices of CDR3α 4mers (left) and CDR3β 4mers (right). PC1 and PC2 are shown. 4mer samples cluster by genotype of the underlying sample.

(C) Number of TCR repertoires (of individuals) in which each unique CDR3α (red), CDR3β (blue) or paired CDR3αβ (purple) motif is observed. 1,696 CDR3α motifs and 644 CDR3β motifs have been observed in TCR repertoires of every single individual analyzed in this study. Identical CDR3αβ motifs have not been observed in more than 12 individuals.

Supplementary Table 1: Detailed sample list of all 32 CITR-seq samples analyzed in the present study.

CITRseq samples using 20h activated T cells							
Sample	**Genotype**	**Sequencing Reads**	**Total Cells with productive TCRs**	**Reads/Cell**	**Total Cells with 1 alpha and 1 beta (paired)**	**Mean UMI/Cell**	**Mean Clones/Cell**
BL6F_1	BL6	33459440	189870	176,22	107400	12,25	2,04
BL6F_2	BL6	62982523	482872	130,43	289492	17,16	2,25
BL6F_4	BL6	41323740	344608	119,92	191474	13,24	2,27
BL6M_3	BL6	62057819	455225	136,32	266110	13,44	2,12
BL6M_4	BL6	49440517	307496	160,78	170893	15,19	2,23
BL6M_5	BL6	41988830	655037	64,10	358823	13,29	2,27
BL6xCastF_1	BL6xCast	51747247	236169	219,11	128310	13,62	2,13
BL6xCastF_2	BL6xCast	49245000	212697	231,53	114385	12,69	2,13
BL6xCastM_2	BL6xCast	49421551	184067	268,50	98821	10,93	2,08
BL6xCastM_3	BL6xCast	57410723	231268	248,24	122758	11,75	2,14
BL6xPwdF_1	BL6xPwd	83185770	319989	259,96	181550	11,69	2,05
BL6xPwdF_2	BL6xPwd	74772578	324570	230,37	178826	11,50	2,06
BL6xPwdM_1	BL6xPwd	44688398	148795	300,34	82313	13,26	2,17
BL6xPwdM_2	BL6xPwd	81521294	320524	254,34	181847	12,26	2,06
BL6xSpretF_1	BL6xSpret	50811359	288939	175,85	165123	13,91	2,08
BL6xSpretF_2	BL6xSpret	35290223	183883	191,92	104313	15,96	2,14
BL6xSpretF_3	BL6xSpret	48325633	347321	139,14	197549	19,09	2,18
BL6xSpretM_1	BL6xSpret	25923260	259761	99,80	151176	24,33	2,18
BL6xSpretM_2	BL6xSpret	49117676	340758	144,14	188112	18,53	2,20
BL6xSpretM_3	BL6xSpret	24096366	239473	100,62	133293	21,74	2,14
CastF_7	Cast	46000265	232486	197,86	135992	15,74	2,01
CastF_8	Cast	38951285	207171	188,02	120875	16,15	2,04
CastM_1	Cast	34737464	191796	181,12	113684	19,08	2,07
CastM_2	Cast	27653742	130700	211,58	74822	15,58	2,02
CastM_3	Cast	55617482	269863	206,10	152614	15,89	2,01
PwdF_1	Pwd	52696502	297699	177,01	151241	12,61	2,08
PwdF_2	Pwd	49311832	328606	150,06	169230	12,72	2,12
PwdM_1	Pwd	84215698	440650	191,12	226863	13,51	2,14
PwdM_2	Pwd	58233463	293925	198,12	154090	12,31	2,06
				184,57		**14,81**	**2,12**
CITRseq samples using Primary T cells							
Sample	**Genotype**	**Sequencing Reads**	**Total Cells with productive TCRs**		**Total Cells with 1 alpha and 1 beta (paired)**	**Mean UMI/Cell**	**Mean Clones/Cell**
SpretF_1	Spret	40464272	275787	146,72	145865	6,52	1,88
SpretM_1	Spret	40418946	158671	254,73	80418	5,34	1,8
SpretM_2	Spret	34073895	212716	160,18	111072	5,96	1,84
				187,21		**5,94**	**1,84**

Supplementary Table 2: List of all DNA oligos and PCR primers used in the present study

T cell receptor alpha V-segment primers			
Target		**Protocol Step**	**Sequence**
Trav1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTTATCCTGGTACCAGCAAC
Trav2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAGGGACCACAGTTTATCATTC
Trav3.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTATCATCTGCACCTACACAGAC
Trav3.2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCCTCACCTGAGTGTCC
Trav4.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCAAGGAACAAAGGAGAATGGAAG
Trav4.2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTCTGTGGGTGCAGATTTGC
Trav4.3	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGGAACAAAGGAGAATGGGAG
Trav5.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAGAGAATCCTAAGCTCATCATTG
Trav5.2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAGCGTCTTCAGTTCATCATAGAC
Trav6.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGCAGAGGTTTTGAAGCTACATATG
Trav6.2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCCTCTCAGAAGAGGACTTTC
Trav6.3	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAGATTCCGTGACTCAAACAG
Trav6.4	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGAAAGCCTCAGTGCAGG
Trav6.5	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAAAGGCCAACGAGAAGGG
Trav7	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGCAGAGCCCAGAATCCC
Trav8.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGTGAAAGTGTCACGGTG
Trav8.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGAGAAGAGAATCTTCAGGC
Trav9.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCCAGTTTCTCCTCAAGTACTATTC
Trav9.2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGTGCTGGGGATACACTTT
Trav9.3	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACCTTATCTGTTCTGGTATGTCCA
Trav10	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAACTGCACTTACACAGATACTGC
Trav11	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGATGCTAAGCACAGCACG
Trav12.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGACAGAAGGCCTGGTC
Trav12.2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGAAGGAAGGCCTGGTC
Trav12.3	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGAACTGCACCTATCAGACT
Trav13.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGCTCTTTGCACATTTCCTC
Trav13.2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCTTTGCACATTTCCTCCTCC
Trav13.3	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGAACGCAGAGCTGCA
Trav13.4	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAGAATGCAGAGCTGCAG
Trav14.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTCTGACAGTCTGGGAAGG
Trav14.2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGGAAGATGGACGATTCACAA
Trav15	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGGCCTTGGCTTTCTCT
Trav16.1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCCTTATTCGTCAGGACTCTTACA
Trav16.2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCTGATTCTAAGCCTGTTGGG
Trav17	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTGGAGCTCAGATGCAG
Trav18	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTTCTGAGTATCCAGGAGGG
Trav19	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCATCGCTGACTGTTCAAGAGG
Trav21	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAATAGTATGGCTTTCCTGGC
Trav23	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCTCTGGTATAGACAAGATCTGG

T cell receptor beta V-segment primers			
Target		**Protocol Step**	**Sequence**
Trb1	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGTATCCCTGGATGAGCTG
Trb2	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTATGGACAATCAGACTGCCTC
Trb3	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATATGGGGCAGATGGTGAC
Trb4	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACGGCTGTTTTCCAGACTC
Trb5	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCAGAGCTCATGTTTCTCT
Trb12	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAGCAGATTCTCAGTCCAAC
Trb13	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGTACTGGTATCGGCAGGAC
Trb14	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCCCAGATATGCAGTCCTAC
Trb15	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTGTGAGCCAGTTTCAGG
Trb16	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAACAAATGCTGGTGTCATCC
Trb17	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGAACAGGGAAGCTGACAC
Trb19	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGGTACCGACAGGATTCAG
Trb20	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGCTTGGTATCGTCAATCGC
Trb21	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAAGAAACCGGGAGAAGAACTC
Trb23	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCAACAGCCTCTTGATCAAATAGAC
Trb26	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATGAGGTGTATCCCTGAAAAGG
Trb29	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTACTGGTATCGACAAGACCC
Trb30	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGACATCTGTCAAAGTGGC
Trb31	Multiplex-PCR		TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTAACCTCTACTGGTACTGGC

T cell receptor constant region primers			
Target		**Protocol Step**	**Sequence**
TCR Constant Alpha	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**AACC**CACAGCAGGTTCTGGGTTCTG
TCR Constant Beta	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**AACC**TTGGGTGGAGTCACATTTCTCAGATC
TCR Constant Alpha	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**GTCA**CACAGCAGGTTCTGGGTTCTG
TCR Constant Beta	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**GTCA**TTGGGTGGAGTCACATTTCTCAGATC
TCR Constant Alpha	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**TGCT**CACAGCAGGTTCTGGGTTCTG
TCR Constant Beta	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**TGCT**TTGGGTGGAGTCACATTTCTCAGATC
TCR Constant Alpha	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**CGAA**CACAGCAGGTTCTGGGTTCTG
TCR Constant Beta	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**CGAA**TTGGGTGGAGTCACATTTCTCAGATC
TCR Constant Alpha	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**ACGA**CACAGCAGGTTCTGGGTTCTG
TCR Constant Beta	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**ACGA**TTGGGTGGAGTCACATTTCTCAGATC
TCR Constant Alpha	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**GAAG**CACAGCAGGTTCTGGGTTCTG
TCR Constant Beta	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**GAAG**TTGGGTGGAGTCACATTTCTCAGATC
TCR Constant Alpha	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**TTGG**CACAGCAGGTTCTGGGTTCTG
TCR Constant Beta	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**TTGG**TTGGGTGGAGTCACATTTCTCAGATC
TCR Constant Alpha	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**AGTG**CACAGCAGGTTCTGGGTTCTG
TCR Constant Beta	Reverse Transcription		/5Phos/GGGCTCGGAGATGTGTATAAGAGACAGNNNNNNNNNNNN**AGTG**TTGGGTGGAGTCACATTTCTCAGATC

Description:
Barcoded RT oligos (8 different barcode pairs for alpha/beta in **bold**). Oligos anneal to the 5' end of TCR alpha and TCR beta constant regions

Round1_081	/5Phos/GTGTGCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGCACACTCT
Round1_082	/5Phos/CATTGCCGTCTCGT	TACACATCTCCGAGCCCACGAGACGGCAATGTCT
Round1_083	/5Phos/CATGTGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCACATGTCT
Round1_084	/5Phos/TGGTAACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTTACCATCT
Round1_085	/5Phos/ATGGAACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTTCCATTCT
Round1_086	/5Phos/AACGACAGTCTCGT	TACACATCTCCGAGCCCACGAGACTGTCGTTTCT
Round1_087	/5Phos/GTTAACCGTCTCGT	TACACATCTCCGAGCCCACGAGACGGTTAACTCT
Round1_088	/5Phos/GAACTGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCAGTTCTCT
Round1_089	/5Phos/CCGATAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTATCGGTCT
Round1_090	/5Phos/AGCGAAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTTCGCTTCT
Round1_091	/5Phos/GAGACGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACGTCTCTCT
Round1_092	/5Phos/AGGCACAGTCTCGT	TACACATCTCCGAGCCCACGAGACTGTGCCTTCT
Round1_093	/5Phos/AATCGGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCCGATTTCT
Round1_094	/5Phos/GGCTCAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTGAGCCTCT
Round1_095	/5Phos/CAGTCAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTGACTGTCT
Round1_096	/5Phos/TGCTGGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCCAGCATCT
Round1_097	/5Phos/AACACAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTGTGTTTCT
Round1_098	/5Phos/AACGTTCGTCTCGT	TACACATCTCCGAGCCCACGAGACGAACGTTTCT
Round1_099	/5Phos/AACTTCGGTCTCGT	TACACATCTCCGAGCCCACGAGACCGAAGTTTCT
Round1_100	/5Phos/AAGCAAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTTGCTTTCT
Round1_101	/5Phos/ACACTCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGAGTGTTCT
Round1_102	/5Phos/ACAGATGGTCTCGT	TACACATCTCCGAGCCCACGAGACCATCTGTTCT
Round1_103	/5Phos/ACATGAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTCATGTTCT
Round1_104	/5Phos/ACGCGATGTCTCGT	TACACATCTCCGAGCCCACGAGACATCGCGTTCT
Round1_105	/5Phos/ACGTCACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTGACGTTCT
Round1_106	/5Phos/ACTAACGGTCTCGT	TACACATCTCCGAGCCCACGAGACCGTTAGTTCT
Round1_107	/5Phos/ACTAGGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCCTAGTTCT
Round1_108	/5Phos/ACTTCCAGTCTCGT	TACACATCTCCGAGCCCACGAGACTGGAAGTTCT
Round1_109	/5Phos/AGAACGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCGTTCTTCT
Round1_110	/5Phos/AGAGCCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGGCTCTTCT
Round1_111	/5Phos/AGCCATCGTCTCGT	TACACATCTCCGAGCCCACGAGACGATGGCTTCT
Round1_112	/5Phos/AGCTCTGGTCTCGT	TACACATCTCCGAGCCCACGAGACCAGAGCTTCT
Round1_113	/5Phos/AGGACAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTGTCCTTCT
Round1_114	/5Phos/AGGTTAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTAACCTTCT
Round1_115	/5Phos/AGTCCATGTCTCGT	TACACATCTCCGAGCCCACGAGACATGGACTTCT
Round1_116	/5Phos/AGTGTCCGTCTCGT	TACACATCTCCGAGCCCACGAGACGGACACTTCT
Round1_117	/5Phos/ATACGGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCCGTATTCT
Round1_118	/5Phos/ATCCTGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCAGGATTCT
Round1_119	/5Phos/ATCTAGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCTAGATTCT
Round1_120	/5Phos/CAACCGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACGGTTGTCT
Round1_121	/5Phos/CAACTCAGTCTCGT	TACACATCTCCGAGCCCACGAGACTGAGTTGTCT
Round1_122	/5Phos/CACATACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTATGTGTCT
Round1_123	/5Phos/CAGAAGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCTTCTGTCT
Round1_124	/5Phos/CAGGACTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGTCCTGTCT
Round1_125	/5Phos/CCAGACAGTCTCGT	TACACATCTCCGAGCCCACGAGACTGTCTGGTCT
Round1_126	/5Phos/CCTAAGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACTTAGGTCT
Round1_127	/5Phos/CGAATCGGTCTCGT	TACACATCTCCGAGCCCACGAGACCGATTCGTCT
Round1_128	/5Phos/CGAGTATGTCTCGT	TACACATCTCCGAGCCCACGAGACATACTCGTCT
Round1_129	/5Phos/CGATCTTGTCTCGT	TACACATCTCCGAGCCCACGAGACAAGATCGTCT
Round1_130	/5Phos/CGTACACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTGTACGTCT
Round1_131	/5Phos/CGTTAAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTTAACGTCT
Round1_132	/5Phos/CGTTCGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCGAACGTCT
Round1_133	/5Phos/CTACTTCGTCTCGT	TACACATCTCCGAGCCCACGAGACGAAGTAGTCT
Round1_134	/5Phos/CTAGGACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTCCTAGTCT
Round1_135	/5Phos/CTATCCGGTCTCGT	TACACATCTCCGAGCCCACGAGACCGGATAGTCT
Round1_136	/5Phos/CTCATTGGTCTCGT	TACACATCTCCGAGCCCACGAGACCAATGAGTCT
Round1_137	/5Phos/CTCTTCAGTCTCGT	TACACATCTCCGAGCCCACGAGACTGAAGAGTCT
Round1_138	/5Phos/CTGATGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCATCAGTCT
Round1_139	/5Phos/CTGGTTAGTCTCGT	TACACATCTCCGAGCCCACGAGACTAACCAGTCT
Round1_140	/5Phos/CTTCTCGGTCTCGT	TACACATCTCCGAGCCCACGAGACCGAGAAGTCT
Round1_141	/5Phos/CTTGCATGTCTCGT	TACACATCTCCGAGCCCACGAGACATGCAAGTCT
Round1_142	/5Phos/GAACCAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTGGTTCTCT
Round1_143	/5Phos/GACCTCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGAGGTCTCT
Round1_144	/5Phos/GAGGATCGTCTCGT	TACACATCTCCGAGCCCACGAGACGATCCTCTCT
Round1_145	/5Phos/GATGGTAGTCTCGT	TACACATCTCCGAGCCCACGAGACTACCATCTCT
Round1_146	/5Phos/GCAAGCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGCTTGCTCT
Round1_147	/5Phos/GCATTGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACAATGCTCT
Round1_148	/5Phos/GCCATAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTATGGCTCT
Round1_149	/5Phos/GCCTAGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCTAGGCTCT
Round1_150	/5Phos/GCTCTGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCAGAGCTCT
Round1_151	/5Phos/GGAGTTGGTCTCGT	TACACATCTCCGAGCCCACGAGACCAACTCCTCT
Round1_152	/5Phos/GGATGACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTCATCCTCT
Round1_153	/5Phos/GGCGATAGTCTCGT	TACACATCTCCGAGCCCACGAGACTATCGCCTCT
Round1_154	/5Phos/GGTCCTAGTCTCGT	TACACATCTCCGAGCCCACGAGACTAGGACCTCT
Round1_155	/5Phos/GTACCTGGTCTCGT	TACACATCTCCGAGCCCACGAGACCAGGTACTCT
Round1_156	/5Phos/GTCATGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACATGACTCT
Round1_157	/5Phos/GTGCAGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCTGCACTCT
Round1_158	/5Phos/GTTGCTCGTCTCGT	TACACATCTCCGAGCCCACGAGACGAGCAACTCT
Round1_159	/5Phos/TACGAACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTTCGTATCT
Round1_160	/5Phos/TAGCTACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTAGCTATCT
Round1_161	/5Phos/TATCGCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGCGATATCT
Round1_162	/5Phos/TCAACAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTGTTGATCT
Round1_163	/5Phos/TCATTCGGTCTCGT	TACACATCTCCGAGCCCACGAGACCGAATGATCT
Round1_164	/5Phos/TCCTTGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCAAGGATCT
Round1_165	/5Phos/TCGACTAGTCTCGT	TACACATCTCCGAGCCCACGAGACTAGTCGATCT
Round1_166	/5Phos/TCGAGACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTCTCGATCT

Blocking Oligo (Blocks the RT overhang before Pooling of cells)		
Target	**Protocol Step**	**Sequence**
RT-oligo overhang	Barcode Ligation	GGGCTCGGAGATGTGTA

Round 1 Barcoding Oligos		
Name	**Bottom Strand**	**Top Strand**
Round1_001	/5Phos/GGTTAGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCTAACCTCT
Round1_002	/5Phos/CACGGAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTCCGTGTCT
Round1_003	/5Phos/GGACAAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTTGTCCTCT
Round1_004	/5Phos/TGGAGCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGCTCCATCT
Round1_005	/5Phos/TCCGTAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTACGGATCT
Round1_006	/5Phos/GATAAGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCTTATCTCT
Round1_007	/5Phos/TGAGCAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTGCTCATCT
Round1_008	/5Phos/ATAGCAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTGCTATTCT
Round1_009	/5Phos/GATCGACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTCGATCTCT
Round1_010	/5Phos/GGAACATGTCTCGT	TACACATCTCCGAGCCCACGAGACATGTTCCTCT
Round1_011	/5Phos/GTAACGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCGTTACTCT
Round1_012	/5Phos/TCTCGTGGTCTCGT	TACACATCTCCGAGCCCACGAGACCACGAGATCT
Round1_013	/5Phos/AGTGGAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTCCACTTCT
Round1_014	/5Phos/GATCACAGTCTCGT	TACACATCTCCGAGCCCACGAGACTGTGATCTCT
Round1_015	/5Phos/TTCGGTAGTCTCGT	TACACATCTCCGAGCCCACGAGACTACCGAATCT
Round1_016	/5Phos/GCCGAATGTCTCGT	TACACATCTCCGAGCCCACGAGACATTCGGCTCT
Round1_017	/5Phos/TAGTGGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCCACTATCT
Round1_018	/5Phos/GAGTTGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCAACTCTCT
Round1_019	/5Phos/CTTACGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCGTAAGTCT
Round1_020	/5Phos/AGAAGCCGTCTCGT	TACACATCTCCGAGCCCACGAGACGGCTTCTTCT
Round1_021	/5Phos/GGTTCCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGGAACCTCT
Round1_022	/5Phos/GCAGGAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTCCTGCTCT
Round1_023	/5Phos/GTGCTATGTCTCGT	TACACATCTCCGAGCCCACGAGACATAGCACTCT
Round1_024	/5Phos/GTGGCAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTGCCACTCT
Round1_025	/5Phos/CGACAGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCTGTCGTCT
Round1_026	/5Phos/AGCTACTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGTAGCTTCT
Round1_027	/5Phos/ACCATCCGTCTCGT	TACACATCTCCGAGCCCACGAGACGGATGGTTCT
Round1_028	/5Phos/ATGCCTAGTCTCGT	TACACATCTCCGAGCCCACGAGACTAGGCATTCT
Round1_029	/5Phos/GTCGTACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTACGACTCT
Round1_030	/5Phos/TACGCCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGGCGTATCT
Round1_031	/5Phos/CTAAGGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACCTTAGTCT
Round1_032	/5Phos/TACGTGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCACGTATCT
Round1_033	/5Phos/CTGACTTGTCTCGT	TACACATCTCCGAGCCCACGAGACAAGTCAGTCT
Round1_034	/5Phos/CGCAATTGTCTCGT	TACACATCTCCGAGCCCACGAGACAATTGCGTCT
Round1_035	/5Phos/GTAGACTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGTCTACTCT
Round1_036	/5Phos/TCCTGCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGCAGGATCT
Round1_037	/5Phos/TTGCCAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTGGCAATCT
Round1_038	/5Phos/TTCAGGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCCTGAATCT
Round1_039	/5Phos/AACAGGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACCTGTTTCT
Round1_040	/5Phos/TTATGCCGTCTCGT	TACACATCTCCGAGCCCACGAGACGGCATAATCT
Round1_041	/5Phos/CGGTATAGTCTCGT	TACACATCTCCGAGCCCACGAGACTATACCGTCT
Round1_042	/5Phos/TTGCATCGTCTCGT	TACACATCTCCGAGCCCACGAGACGATGCAATCT
Round1_043	/5Phos/CATACCTGTCTCGT	TACACATCTCCGAGCCCACGAGACAGGTATGTCT
Round1_044	/5Phos/ATGGTGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACACCATTCT
Round1_045	/5Phos/GCTTATGGTCTCGT	TACACATCTCCGAGCCCACGAGACCATAAGCTCT
Round1_046	/5Phos/TAATCGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCGATTATCT
Round1_047	/5Phos/GTATAGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCTATACTCT
Round1_048	/5Phos/CCTTCTCGTCTCGT	TACACATCTCCGAGCCCACGAGACGAGAAGGTCT
Round1_049	/5Phos/CCACTAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTAGTGGTCT
Round1_050	/5Phos/AATGAGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCTCATTTCT
Round1_051	/5Phos/GTGATCGGTCTCGT	TACACATCTCCGAGCCCACGAGACCGATCACTCT
Round1_052	/5Phos/ACCAGAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTCTGGTTCT
Round1_053	/5Phos/TGACTGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACAGTCATCT
Round1_054	/5Phos/TGGTCCAGTCTCGT	TACACATCTCCGAGCCCACGAGACTGGACCATCT
Round1_055	/5Phos/CCGAATCGTCTCGT	TACACATCTCCGAGCCCACGAGACGATTCGGTCT
Round1_056	/5Phos/ACACCGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCGGTGTTCT
Round1_057	/5Phos/TGTGGTTGTCTCGT	TACACATCTCCGAGCCCACGAGACAACCACATCT
Round1_058	/5Phos/TCAAGGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCCTTGATCT
Round1_059	/5Phos/GGTAGGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCCTACCTCT
Round1_060	/5Phos/GAAGAGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCTCTTCTCT
Round1_061	/5Phos/ATGTCGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCGACATTCT
Round1_062	/5Phos/TATACGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCGTATATCT
Round1_063	/5Phos/GCCTCTTGTCTCGT	TACACATCTCCGAGCCCACGAGACAAGAGGCTCT
Round1_064	/5Phos/TGTGAGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCTCACATCT
Round1_065	/5Phos/AACCGTGGTCTCGT	TACACATCTCCGAGCCCACGAGACCACGGTTTCT
Round1_066	/5Phos/CAGCGTAGTCTCGT	TACACATCTCCGAGCCCACGAGACTACGCTGTCT
Round1_067	/5Phos/TCGTAGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACTACGATCT
Round1_068	/5Phos/ATCGGATGTCTCGT	TACACATCTCCGAGCCCACGAGACATCCGATTCT
Round1_069	/5Phos/GGAATGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCATTCCTCT
Round1_070	/5Phos/AATGGCGGTCTCGT	TACACATCTCCGAGCCCACGAGACCGCCATTTCT
Round1_071	/5Phos/CTGTGAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTCACAGTCT
Round1_072	/5Phos/CGAAGAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTCTTCGTCT
Round1_073	/5Phos/TCTCCACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTGGAGATCT
Round1_074	/5Phos/CCTCGAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTCGAGGTCT
Round1_075	/5Phos/AGGAATGGTCTCGT	TACACATCTCCGAGCCCACGAGACCATTCCTTCT
Round1_076	/5Phos/CTCACAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTGTGAGTCT
Round1_077	/5Phos/ATGAGGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCCTCATTCT
Round1_078	/5Phos/CTGTACCGTCTCGT	TACACATCTCCGAGCCCACGAGACGGTACAGTCT
Round1_079	/5Phos/TGCACGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACGTGCATCT
Round1_080	/5Phos/CCTGAACGTCTCGT	TACACATCTCCGAGCCCACGAGACGTTCAGGTCT

Round1_167	/5Phos/TCGGAAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTTCCGATCT
Round1_168	/5Phos/TCGGTTCGTCTCGT	TACACATCTCCGAGCCCACGAGACGAACCGATCT
Round1_169	/5Phos/TCTTACCGTCTCGT	TACACATCTCCGAGCCCACGAGACGGTAAGATCT
Round1_170	/5Phos/TGAGACGGTCTCGT	TACACATCTCCGAGCCCACGAGACCGTCTCATCT
Round1_171	/5Phos/TGCAACCGTCTCGT	TACACATCTCCGAGCCCACGAGACGGTTGCATCT
Round1_172	/5Phos/TGCAGAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTCTGCATCT
Round1_173	/5Phos/TGGATGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCATCCATCT
Round1_174	/5Phos/TGGCGAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTCGCCATCT
Round1_175	/5Phos/TGTCTAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTAGACATCT
Round1_176	/5Phos/TTAGAGCGTCTCGT	TACACATCTCCGAGCCCACGAGACGCTCTAATCT
Round1_177	/5Phos/TTGAAGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCTTCAATCT
Round1_178	/5Phos/TTGCGGTGTCTCGT	TACACATCTCCGAGCCCACGAGACACCGCAATCT
Round1_179	/5Phos/CTCCGTTGTCTCGT	TACACATCTCCGAGCCCACGAGACAACGGAGTCT
Round1_180	/5Phos/AGCAAGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCTTGCTTCT
Round1_181	/5Phos/GAGAGAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTCTCTCTCT
Round1_182	/5Phos/GCTTGATGTCTCGT	TACACATCTCCGAGCCCACGAGACATCAAGCTCT
Round1_183	/5Phos/ATTGCGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCGCAATTCT
Round1_184	/5Phos/GACACTAGTCTCGT	TACACATCTCCGAGCCCACGAGACTAGTGTCTCT
Round1_185	/5Phos/AAGATGGGTCTCGT	TACACATCTCCGAGCCCACGAGACCCATCTTTCT
Round1_186	/5Phos/GAGGTAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTACCTCTCT
Round1_187	/5Phos/CCATCAAGTCTCGT	TACACATCTCCGAGCCCACGAGACTTGATGGTCT
Round1_188	/5Phos/ACGCTTGGTCTCGT	TACACATCTCCGAGCCCACGAGACCAAGCGTTCT
Round1_189	/5Phos/ATTCGAGGTCTCGT	TACACATCTCCGAGCCCACGAGACCTCGAATTCT
Round1_190	/5Phos/TGACCTCGTCTCGT	TACACATCTCCGAGCCCACGAGACGAGGTCATCT
Round1_191	/5Phos/ACGGAGAGTCTCGT	TACACATCTCCGAGCCCACGAGACTCTCCGTTCT
Round1_192	/5Phos/AAGGCTGGTCTCGT	TACACATCTCCGAGCCCACGAGACCAGCCTTTCT

Description:

#Round 1 barcode oligos

#Each barcode consists of one bottom strand oligo + it's matching top strand oligo which together will form the duplex

(rev. comp. part for most of the sequence, leaving some single stranded overhangs used for ligation)

#Round 1 oligos need to be phosphorylated in order to ligate the round 2 barcodes

#Round 1 barcodes are positioned in bases 1-7 of index 1, the actual barcode are the first seven bases of the oligo sequence

Round 2 Barcoding Oligos		
Name	**Bottom Strand**	**Top Strand**
Round2_001	CAAGCAGAAGACGGCATACGAGATTCTGGCAAGA	TGCCAGAATCTCGTATGCCGTCTTCTGCTTG
Round2_002	CAAGCAGAAGACGGCATACGAGATGAACGTTAGA	AACGTTCATCTCGTATGCCGTCTTCTGCTTG
Round2_003	CAAGCAGAAGACGGCATACGAGATGAATCTCAGA	GAGATTCATCTCGTATGCCGTCTTCTGCTTG
Round2_004	CAAGCAGAAGACGGCATACGAGATAGCAAGAAGA	TCTTGCTATCTCGTATGCCGTCTTCTGCTTG
Round2_005	CAAGCAGAAGACGGCATACGAGATACGCTTGAGA	CAAGCGTATCTCGTATGCCGTCTTCTGCTTG
Round2_006	CAAGCAGAAGACGGCATACGAGATGACTAAGAGA	CTTAGTCATCTCGTATGCCGTCTTCTGCTTG
Round2_007	CAAGCAGAAGACGGCATACGAGATTCACAGGAGA	CCTGTGAATCTCGTATGCCGTCTTCTGCTTG
Round2_008	CAAGCAGAAGACGGCATACGAGATGTTCCGTAGA	ACGGAACATCTCGTATGCCGTCTTCTGCTTG
Round2_009	CAAGCAGAAGACGGCATACGAGATGGTGTGTAGA	ACACACCATCTCGTATGCCGTCTTCTGCTTG
Round2_010	CAAGCAGAAGACGGCATACGAGATTCAGCGTAGA	ACGCTGAATCTCGTATGCCGTCTTCTGCTTG
Round2_011	CAAGCAGAAGACGGCATACGAGATTAGAGTGAGA	CACTCTAATCTCGTATGCCGTCTTCTGCTTG
Round2_012	CAAGCAGAAGACGGCATACGAGATCCAAGTGAGA	CACTTGGATCTCGTATGCCGTCTTCTGCTTG
Round2_013	CAAGCAGAAGACGGCATACGAGATAAGTGCAAGA	TGCACTTATCTCGTATGCCGTCTTCTGCTTG
Round2_014	CAAGCAGAAGACGGCATACGAGATTAAGGTCAGA	GACCTTAATCTCGTATGCCGTCTTCTGCTTG
Round2_015	CAAGCAGAAGACGGCATACGAGATATGCTCCAGA	GGAGCATATCTCGTATGCCGTCTTCTGCTTG
Round2_016	CAAGCAGAAGACGGCATACGAGATCTCCAAGAGA	CTTGGAGATCTCGTATGCCGTCTTCTGCTTG
Round2_017	CAAGCAGAAGACGGCATACGAGATAAGATGGAGA	CCATCTTATCTCGTATGCCGTCTTCTGCTTG
Round2_018	CAAGCAGAAGACGGCATACGAGATAGCCTAAAGA	TTAGGCTATCTCGTATGCCGTCTTCTGCTTG
Round2_019	CAAGCAGAAGACGGCATACGAGATGGATACAAGA	TGTATCCATCTCGTATGCCGTCTTCTGCTTG
Round2_020	CAAGCAGAAGACGGCATACGAGATGTTGAAGAGA	CTTCAACATCTCGTATGCCGTCTTCTGCTTG
Round2_021	CAAGCAGAAGACGGCATACGAGATCGCTGATAGA	ATCAGCGATCTCGTATGCCGTCTTCTGCTTG
Round2_022	CAAGCAGAAGACGGCATACGAGATTACAGCAAGA	TGCTGTAATCTCGTATGCCGTCTTCTGCTTG
Round2_023	CAAGCAGAAGACGGCATACGAGATCGAGATCAGA	GATCTCGATCTCGTATGCCGTCTTCTGCTTG
Round2_024	CAAGCAGAAGACGGCATACGAGATACACAACAGA	GTTGTGTATCTCGTATGCCGTCTTCTGCTTG
Round2_025	CAAGCAGAAGACGGCATACGAGATTTCTCACAGA	GTGAGAAATCTCGTATGCCGTCTTCTGCTTG
Round2_026	CAAGCAGAAGACGGCATACGAGATGCGTGTAAGA	TACACGCATCTCGTATGCCGTCTTCTGCTTG
Round2_027	CAAGCAGAAGACGGCATACGAGATACGGAGAAGA	TCTCCGTATCTCGTATGCCGTCTTCTGCTTG
Round2_028	CAAGCAGAAGACGGCATACGAGATCTAGTGGAGA	CCACTAGATCTCGTATGCCGTCTTCTGCTTG
Round2_029	CAAGCAGAAGACGGCATACGAGATATTGCGAAGA	TCGCAATATCTCGTATGCCGTCTTCTGCTTG
Round2_030	CAAGCAGAAGACGGCATACGAGATTTGGACAAGA	TGTCCAAATCTCGTATGCCGTCTTCTGCTTG
Round2_031	CAAGCAGAAGACGGCATACGAGATATTCGAGAGA	CTCGAATATCTCGTATGCCGTCTTCTGCTTG
Round2_032	CAAGCAGAAGACGGCATACGAGATAACTGACAGA	GTCAGTTATCTCGTATGCCGTCTTCTGCTTG
Round2_033	CAAGCAGAAGACGGCATACGAGATAGTCAGGAGA	CCTGACTATCTCGTATGCCGTCTTCTGCTTG
Round2_034	CAAGCAGAAGACGGCATACGAGATACGACCTAGA	AGGTCGTATCTCGTATGCCGTCTTCTGCTTG
Round2_035	CAAGCAGAAGACGGCATACGAGATTGACCTCAGA	GAGGTCAATCTCGTATGCCGTCTTCTGCTTG
Round2_036	CAAGCAGAAGACGGCATACGAGATCCTGTCTAGA	AGACAGGATCTCGTATGCCGTCTTCTGCTTG
Round2_037	CAAGCAGAAGACGGCATACGAGATTGTTGCGAGA	CGCAACAATCTCGTATGCCGTCTTCTGCTTG
Round2_038	CAAGCAGAAGACGGCATACGAGATCGGTTGTAGA	ACAACCGATCTCGTATGCCGTCTTCTGCTTG
Round2_039	CAAGCAGAAGACGGCATACGAGATCAAGAAGAGA	CTTCTTGATCTCGTATGCCGTCTTCTGCTTG
Round2_040	CAAGCAGAAGACGGCATACGAGATGCTATTCAGA	GAATAGCATCTCGTATGCCGTCTTCTGCTTG
Round2_041	CAAGCAGAAGACGGCATACGAGATTAGGCGAAGA	TCGCCTAATCTCGTATGCCGTCTTCTGCTTG
Round2_042	CAAGCAGAAGACGGCATACGAGATGTTGTCAAGA	TGACAACATCTCGTATGCCGTCTTCTGCTTG
Round2_043	CAAGCAGAAGACGGCATACGAGATAATCCTCAGA	GAGGATTATCTCGTATGCCGTCTTCTGCTTG
Round2_044	CAAGCAGAAGACGGCATACGAGATTAGTACGAGA	CGTACTAATCTCGTATGCCGTCTTCTGCTTG
Round2_045	CAAGCAGAAGACGGCATACGAGATACCGCTAAGA	TAGCGGTATCTCGTATGCCGTCTTCTGCTTG
Round2_046	CAAGCAGAAGACGGCATACGAGATCATAGAGAGA	CTCTATGATCTCGTATGCCGTCTTCTGCTTG
Round2_047	CAAGCAGAAGACGGCATACGAGATCTTGGTGAGA	CACCAAGATCTCGTATGCCGTCTTCTGCTTG
Round2_048	CAAGCAGAAGACGGCATACGAGATATACACGAGA	CGTGTATATCTCGTATGCCGTCTTCTGCTTG
Round2_049	CAAGCAGAAGACGGCATACGAGATCACAACGAGA	CGTTGTGATCTCGTATGCCGTCTTCTGCTTG
Round2_050	CAAGCAGAAGACGGCATACGAGATATAGGCAAGA	TGCCTATATCTCGTATGCCGTCTTCTGCTTG

Round2_051	CAAGCAGAAGACGGCATACGAGATAGACGTAAGA	TACGTCTATCTCGTATGCCGTCTTCTGCTTG
Round2_052	CAAGCAGAAGACGGCATACGAGATTTCCACTAGA	AGTGGAAATCTCGTATGCCGTCTTCTGCTTG
Round2_053	CAAGCAGAAGACGGCATACGAGATTATGCAGAGA	CTGCATAATCTCGTATGCCGTCTTCTGCTTG
Round2_054	CAAGCAGAAGACGGCATACGAGATAGATGGTAGA	ACCATCTATCTCGTATGCCGTCTTCTGCTTG
Round2_055	CAAGCAGAAGACGGCATACGAGATCTTCAGCAGA	GCTGAAGATCTCGTATGCCGTCTTCTGCTTG
Round2_056	CAAGCAGAAGACGGCATACGAGATCTCCGTTAGA	AACGGAGATCTCGTATGCCGTCTTCTGCTTG
Round2_057	CAAGCAGAAGACGGCATACGAGATGTCTGTGAGA	CACAGACATCTCGTATGCCGTCTTCTGCTTG
Round2_058	CAAGCAGAAGACGGCATACGAGATCCATCAAAGA	TTGATGGATCTCGTATGCCGTCTTCTGCTTG
Round2_059	CAAGCAGAAGACGGCATACGAGATGAGGTAGAGA	CTACCTCATCTCGTATGCCGTCTTCTGCTTG
Round2_060	CAAGCAGAAGACGGCATACGAGATTGTACTGAGA	CAGTACAATCTCGTATGCCGTCTTCTGCTTG
Round2_061	CAAGCAGAAGACGGCATACGAGATAGAGTGAAGA	TCACTCTATCTCGTATGCCGTCTTCTGCTTG
Round2_062	CAAGCAGAAGACGGCATACGAGATTAACACCAGA	GGTGTTAATCTCGTATGCCGTCTTCTGCTTG
Round2_063	CAAGCAGAAGACGGCATACGAGATATCAGCGAGA	CGCTGATATCTCGTATGCCGTCTTCTGCTTG
Round2_064	CAAGCAGAAGACGGCATACGAGATAAGAACCAGA	GGTTCTTATCTCGTATGCCGTCTTCTGCTTG
Round2_065	CAAGCAGAAGACGGCATACGAGATAAGGCTGAGA	CAGCCTTATCTCGTATGCCGTCTTCTGCTTG
Round2_066	CAAGCAGAAGACGGCATACGAGATCTCTCGTAGA	ACGAGAGATCTCGTATGCCGTCTTCTGCTTG
Round2_067	CAAGCAGAAGACGGCATACGAGATGAAGTCCAGA	GGACTTCATCTCGTATGCCGTCTTCTGCTTG
Round2_068	CAAGCAGAAGACGGCATACGAGATAGGTGTCAGA	GACACCTATCTCGTATGCCGTCTTCTGCTTG
Round2_069	CAAGCAGAAGACGGCATACGAGATGCACATAAGA	TATGTGCATCTCGTATGCCGTCTTCTGCTTG
Round2_070	CAAGCAGAAGACGGCATACGAGATCTCGAGAAGA	TCTCGAGATCTCGTATGCCGTCTTCTGCTTG
Round2_071	CAAGCAGAAGACGGCATACGAGATCGTAACAAGA	TGTTACGATCTCGTATGCCGTCTTCTGCTTG
Round2_072	CAAGCAGAAGACGGCATACGAGATGTAAGAGAGA	CTCTTACATCTCGTATGCCGTCTTCTGCTTG
Round2_073	CAAGCAGAAGACGGCATACGAGATAACTCGAAGA	TCGAGTTATCTCGTATGCCGTCTTCTGCTTG
Round2_074	CAAGCAGAAGACGGCATACGAGATGAGAGAAAGA	TTCTCTCATCTCGTATGCCGTCTTCTGCTTG
Round2_075	CAAGCAGAAGACGGCATACGAGATACAACTCAGA	GAGTTGTATCTCGTATGCCGTCTTCTGCTTG
Round2_076	CAAGCAGAAGACGGCATACGAGATCGGCAATAGA	ATTGCCGATCTCGTATGCCGTCTTCTGCTTG
Round2_077	CAAGCAGAAGACGGCATACGAGATCAAGCTAAGA	TAGCTTGATCTCGTATGCCGTCTTCTGCTTG
Round2_078	CAAGCAGAAGACGGCATACGAGATGAATGCGAGA	CGCATTCATCTCGTATGCCGTCTTCTGCTTG
Round2_079	CAAGCAGAAGACGGCATACGAGATTCGCTCAAGA	TGAGCGAATCTCGTATGCCGTCTTCTGCTTG
Round2_080	CAAGCAGAAGACGGCATACGAGATTCTATGGAGA	CCATAGAATCTCGTATGCCGTCTTCTGCTTG
Round2_081	CAAGCAGAAGACGGCATACGAGATCCATAGCAGA	GCTATGGATCTCGTATGCCGTCTTCTGCTTG
Round2_082	CAAGCAGAAGACGGCATACGAGATGCTACAAAGA	TTGTAGCATCTCGTATGCCGTCTTCTGCTTG
Round2_083	CAAGCAGAAGACGGCATACGAGATGTCCGAAAGA	TTCGGACATCTCGTATGCCGTCTTCTGCTTG
Round2_084	CAAGCAGAAGACGGCATACGAGATTCCAATGAGA	CATTGGAATCTCGTATGCCGTCTTCTGCTTG
Round2_085	CAAGCAGAAGACGGCATACGAGATCACTATCAGA	GATAGTGATCTCGTATGCCGTCTTCTGCTTG
Round2_086	CAAGCAGAAGACGGCATACGAGATCAATGGAAGA	TCCATTGATCTCGTATGCCGTCTTCTGCTTG
Round2_087	CAAGCAGAAGACGGCATACGAGATTACCAGAAGA	TCTGGTAATCTCGTATGCCGTCTTCTGCTTG
Round2_088	CAAGCAGAAGACGGCATACGAGATGCGAACAAGA	TGTTCGCATCTCGTATGCCGTCTTCTGCTTG
Round2_089	CAAGCAGAAGACGGCATACGAGATTAACGAGAGA	CTCGTTAATCTCGTATGCCGTCTTCTGCTTG
Round2_090	CAAGCAGAAGACGGCATACGAGATGCTTGATAGA	ATCAAGCATCTCGTATGCCGTCTTCTGCTTG
Round2_091	CAAGCAGAAGACGGCATACGAGATTAGCCTTAGA	AAGGCTAATCTCGTATGCCGTCTTCTGCTTG
Round2_092	CAAGCAGAAGACGGCATACGAGATGTGAGTCAGA	GACTCACATCTCGTATGCCGTCTTCTGCTTG
Round2_093	CAAGCAGAAGACGGCATACGAGATACTGTAGAGA	CTACAGTATCTCGTATGCCGTCTTCTGCTTG
Round2_094	CAAGCAGAAGACGGCATACGAGATCATCATGAGA	CATGATGATCTCGTATGCCGTCTTCTGCTTG
Round2_095	CAAGCAGAAGACGGCATACGAGATGACACTAAGA	TAGTGTCATCTCGTATGCCGTCTTCTGCTTG
Round2_096	CAAGCAGAAGACGGCATACGAGATGGCATCAAGA	TGATGCCATCTCGTATGCCGTCTTCTGCTTG
Round2_097	CAAGCAGAAGACGGCATACGAGATAGGACAAAGA	TTGTCCTATCTCGTATGCCGTCTTCTGCTTG
Round2_098	CAAGCAGAAGACGGCATACGAGATGTCGTACAGA	GTACGACATCTCGTATGCCGTCTTCTGCTTG
Round2_099	CAAGCAGAAGACGGCATACGAGATGAGACGTAGA	ACGTCTCATCTCGTATGCCGTCTTCTGCTTG
Round2_100	CAAGCAGAAGACGGCATACGAGATTAGTGGCAGA	GCCACTAATCTCGTATGCCGTCTTCTGCTTG
Round2_101	CAAGCAGAAGACGGCATACGAGATGTGGCAAAGA	TTGCCACATCTCGTATGCCGTCTTCTGCTTG
Round2_102	CAAGCAGAAGACGGCATACGAGATTTGCATCAGA	GATGCAAATCTCGTATGCCGTCTTCTGCTTG
Round2_103	CAAGCAGAAGACGGCATACGAGATCAGAAGAAGA	TCTTCTGATCTCGTATGCCGTCTTCTGCTTG
Round2_104	CAAGCAGAAGACGGCATACGAGATAGAGCCTAGA	AGGCTCTATCTCGTATGCCGTCTTCTGCTTG
Round2_105	CAAGCAGAAGACGGCATACGAGATGTACCTGAGA	CAGGTACATCTCGTATGCCGTCTTCTGCTTG
Round2_106	CAAGCAGAAGACGGCATACGAGATTCCTTGCAGA	GCAAGGAATCTCGTATGCCGTCTTCTGCTTG
Round2_107	CAAGCAGAAGACGGCATACGAGATCGATCTTAGA	AAGATCGATCTCGTATGCCGTCTTCTGCTTG
Round2_108	CAAGCAGAAGACGGCATACGAGATCACGGAAAGA	TTCCGTGATCTCGTATGCCGTCTTCTGCTTG
Round2_109	CAAGCAGAAGACGGCATACGAGATTGCACGTAGA	ACGTGCAATCTCGTATGCCGTCTTCTGCTTG
Round2_110	CAAGCAGAAGACGGCATACGAGATAGGTTAGAGA	CTAACCTATCTCGTATGCCGTCTTCTGCTTG
Round2_111	CAAGCAGAAGACGGCATACGAGATCGCAATTAGA	AATTGCGATCTCGTATGCCGTCTTCTGCTTG
Round2_112	CAAGCAGAAGACGGCATACGAGATCAGTCAGAGA	CTGACTGATCTCGTATGCCGTCTTCTGCTTG
Round2_113	CAAGCAGAAGACGGCATACGAGATAGTGGAAAGA	TTCCACTATCTCGTATGCCGTCTTCTGCTTG
Round2_114	CAAGCAGAAGACGGCATACGAGATAACACAGAGA	CTGTGTTATCTCGTATGCCGTCTTCTGCTTG
Round2_115	CAAGCAGAAGACGGCATACGAGATCGAATCGAGA	CGATTCGATCTCGTATGCCGTCTTCTGCTTG
Round2_116	CAAGCAGAAGACGGCATACGAGATTGGCGAAAGA	TTCGCCAATCTCGTATGCCGTCTTCTGCTTG
Round2_117	CAAGCAGAAGACGGCATACGAGATATGGAACAGA	GTTCCATATCTCGTATGCCGTCTTCTGCTTG
Round2_118	CAAGCAGAAGACGGCATACGAGATCCGATAAAGA	TTATCGGATCTCGTATGCCGTCTTCTGCTTG
Round2_119	CAAGCAGAAGACGGCATACGAGATGTAGACTAGA	AGTCTACATCTCGTATGCCGTCTTCTGCTTG
Round2_120	CAAGCAGAAGACGGCATACGAGATGAGGATCAGA	GATCCTCATCTCGTATGCCGTCTTCTGCTTG
Round2_121	CAAGCAGAAGACGGCATACGAGATGCATTGTAGA	ACAATGCATCTCGTATGCCGTCTTCTGCTTG
Round2_122	CAAGCAGAAGACGGCATACGAGATCCGAATCAGA	GATTCGGATCTCGTATGCCGTCTTCTGCTTG
Round2_123	CAAGCAGAAGACGGCATACGAGATACCATCCAGA	GGATGGTATCTCGTATGCCGTCTTCTGCTTG
Round2_124	CAAGCAGAAGACGGCATACGAGATACTAACGAGA	CGTTAGTATCTCGTATGCCGTCTTCTGCTTG
Round2_125	CAAGCAGAAGACGGCATACGAGATACTAGGCAGA	GCCTAGTATCTCGTATGCCGTCTTCTGCTTG
Round2_126	CAAGCAGAAGACGGCATACGAGATTCGAGACAGA	GTCTCGAATCTCGTATGCCGTCTTCTGCTTG
Round2_127	CAAGCAGAAGACGGCATACGAGATATCCTGGAGA	CCAGGATATCTCGTATGCCGTCTTCTGCTTG
Round2_128	CAAGCAGAAGACGGCATACGAGATAGCGAAGAGA	CTTCGCTATCTCGTATGCCGTCTTCTGCTTG
Round2_129	CAAGCAGAAGACGGCATACGAGATGTTGCTCAGA	GAGCAACATCTCGTATGCCGTCTTCTGCTTG
Round2_130	CAAGCAGAAGACGGCATACGAGATCATTGCCAGA	GGCAATGATCTCGTATGCCGTCTTCTGCTTG
Round2_131	CAAGCAGAAGACGGCATACGAGATGCTCTGAAGA	TCAGAGCATCTCGTATGCCGTCTTCTGCTTG
Round2_132	CAAGCAGAAGACGGCATACGAGATCACATACAGA	GTATGTGATCTCGTATGCCGTCTTCTGCTTG
Round2_133	CAAGCAGAAGACGGCATACGAGATCCACTAGAGA	CTAGTGGATCTCGTATGCCGTCTTCTGCTTG
Round2_134	CAAGCAGAAGACGGCATACGAGATTCCGTAAAGA	TTACGGAATCTCGTATGCCGTCTTCTGCTTG
Round2_135	CAAGCAGAAGACGGCATACGAGATCCTTCTCAGA	GAGAAGGATCTCGTATGCCGTCTTCTGCTTG
Round2_136	CAAGCAGAAGACGGCATACGAGATGTCATGTAGA	ACATGACATCTCGTATGCCGTCTTCTGCTTG

Round2_137	CAAGCAGAAGACGGCATACGAGATTTGCGGTAGA	ACCGCAAATCTCGTATGCCGTCTTCTGCTTG
Round2_138	CAAGCAGAAGACGGCATACGAGATGGTAGGAAGA	TCCTACCATCTCGTATGCCGTCTTCTGCTTG
Round2_139	CAAGCAGAAGACGGCATACGAGATGGAGTTGAGA	CAACTCCATCTCGTATGCCGTCTTCTGCTTG
Round2_140	CAAGCAGAAGACGGCATACGAGATCTGTGAAAGA	TTCACAGATCTCGTATGCCGTCTTCTGCTTG
Round2_141	CAAGCAGAAGACGGCATACGAGATTGCTGGAAGA	TCCAGCAATCTCGTATGCCGTCTTCTGCTTG
Round2_142	CAAGCAGAAGACGGCATACGAGATACACCGAAGA	TCGGTGTATCTCGTATGCCGTCTTCTGCTTG
Round2_143	CAAGCAGAAGACGGCATACGAGATCGACAGAAGA	TCTGTCGATCTCGTATGCCGTCTTCTGCTTG
Round2_144	CAAGCAGAAGACGGCATACGAGATCTTGCATAGA	ATGCAAGATCTCGTATGCCGTCTTCTGCTTG
Round2_145	CAAGCAGAAGACGGCATACGAGATCCTAAGTAGA	ACTTAGGATCTCGTATGCCGTCTTCTGCTTG
Round2_146	CAAGCAGAAGACGGCATACGAGATTCTCCACAGA	GTGGAGAATCTCGTATGCCGTCTTCTGCTTG
Round2_147	CAAGCAGAAGACGGCATACGAGATTGACTGTAGA	ACAGTCAATCTCGTATGCCGTCTTCTGCTTG
Round2_148	CAAGCAGAAGACGGCATACGAGATCGTTCGAAGA	TCGAACGATCTCGTATGCCGTCTTCTGCTTG
Round2_149	CAAGCAGAAGACGGCATACGAGATTGTGAGAAGA	TCTCACAATCTCGTATGCCGTCTTCTGCTTG
Round2_150	CAAGCAGAAGACGGCATACGAGATCGTACACAGA	GTGTACGATCTCGTATGCCGTCTTCTGCTTG
Round2_151	CAAGCAGAAGACGGCATACGAGATTGTGGTTAGA	AACCACAATCTCGTATGCCGTCTTCTGCTTG
Round2_152	CAAGCAGAAGACGGCATACGAGATATCTAGCAGA	GCTAGATATCTCGTATGCCGTCTTCTGCTTG
Round2_153	CAAGCAGAAGACGGCATACGAGATGTAACGAAGA	TCGTTACATCTCGTATGCCGTCTTCTGCTTG
Round2_154	CAAGCAGAAGACGGCATACGAGATCAGCGTAAGA	TACGCTGATCTCGTATGCCGTCTTCTGCTTG
Round2_155	CAAGCAGAAGACGGCATACGAGATGTATAGGAGA	CCTATACATCTCGTATGCCGTCTTCTGCTTG
Round2_156	CAAGCAGAAGACGGCATACGAGATCTAAGGTAGA	ACCTTAGATCTCGTATGCCGTCTTCTGCTTG
Round2_157	CAAGCAGAAGACGGCATACGAGATCGGTATAAGA	TATACCGATCTCGTATGCCGTCTTCTGCTTG
Round2_158	CAAGCAGAAGACGGCATACGAGATACACTCTAGA	AGAGTGTATCTCGTATGCCGTCTTCTGCTTG
Round2_159	CAAGCAGAAGACGGCATACGAGATGCCATAGAGA	CTATGGCATCTCGTATGCCGTCTTCTGCTTG
Round2_160	CAAGCAGAAGACGGCATACGAGATAGAACGGAGA	CCGTTCTATCTCGTATGCCGTCTTCTGCTTG
Round2_161	CAAGCAGAAGACGGCATACGAGATTACGAACAGA	GTTCGTAATCTCGTATGCCGTCTTCTGCTTG
Round2_162	CAAGCAGAAGACGGCATACGAGATGTGCAGAAGA	TCTGCACATCTCGTATGCCGTCTTCTGCTTG
Round2_163	CAAGCAGAAGACGGCATACGAGATAGCTCTGAGA	CAGAGCTATCTCGTATGCCGTCTTCTGCTTG
Round2_164	CAAGCAGAAGACGGCATACGAGATGGTTCCTAGA	AGGAACCATCTCGTATGCCGTCTTCTGCTTG
Round2_165	CAAGCAGAAGACGGCATACGAGATCCTCGAAAGA	TTCGAGGATCTCGTATGCCGTCTTCTGCTTG
Round2_166	CAAGCAGAAGACGGCATACGAGATGATCACAAGA	TGTGATCATCTCGTATGCCGTCTTCTGCTTG
Round2_167	CAAGCAGAAGACGGCATACGAGATCGAGTATAGA	ATACTCGATCTCGTATGCCGTCTTCTGCTTG
Round2_168	CAAGCAGAAGACGGCATACGAGATCTGGTTAAGA	TAACCAGATCTCGTATGCCGTCTTCTGCTTG
Round2_169	CAAGCAGAAGACGGCATACGAGATGATAAGGAGA	CCTTATCATCTCGTATGCCGTCTTCTGCTTG
Round2_170	CAAGCAGAAGACGGCATACGAGATGCTTATGAGA	CATAAGCATCTCGTATGCCGTCTTCTGCTTG
Round2_171	CAAGCAGAAGACGGCATACGAGATCTAGGACAGA	GTCCTAGATCTCGTATGCCGTCTTCTGCTTG
Round2_172	CAAGCAGAAGACGGCATACGAGATACATGAGAGA	CTCATGTATCTCGTATGCCGTCTTCTGCTTG
Round2_173	CAAGCAGAAGACGGCATACGAGATATGTCGGAGA	CCGACATATCTCGTATGCCGTCTTCTGCTTG
Round2_174	CAAGCAGAAGACGGCATACGAGATTACGTGGAGA	CCACGTAATCTCGTATGCCGTCTTCTGCTTG
Round2_175	CAAGCAGAAGACGGCATACGAGATACGCGATAGA	ATCGCGTATCTCGTATGCCGTCTTCTGCTTG
Round2_176	CAAGCAGAAGACGGCATACGAGATATCGGATAGA	ATCCGATATCTCGTATGCCGTCTTCTGCTTG
Round2_177	CAAGCAGAAGACGGCATACGAGATAACAGGTAGA	ACCTGTTATCTCGTATGCCGTCTTCTGCTTG
Round2_178	CAAGCAGAAGACGGCATACGAGATTCGGTTCAGA	GAACCGAATCTCGTATGCCGTCTTCTGCTTG
Round2_179	CAAGCAGAAGACGGCATACGAGATCTACTTCAGA	GAAGTAGATCTCGTATGCCGTCTTCTGCTTG
Round2_180	CAAGCAGAAGACGGCATACGAGATCTGATGCAGA	GCATCAGATCTCGTATGCCGTCTTCTGCTTG
Round2_181	CAAGCAGAAGACGGCATACGAGATTCTCGTGAGA	CACGAGAATCTCGTATGCCGTCTTCTGCTTG
Round2_182	CAAGCAGAAGACGGCATACGAGATTGCAACCAGA	GGTTGCAATCTCGTATGCCGTCTTCTGCTTG
Round2_183	CAAGCAGAAGACGGCATACGAGATTGAGCAAAGA	TTGCTCAATCTCGTATGCCGTCTTCTGCTTG
Round2_184	CAAGCAGAAGACGGCATACGAGATTGAGACGAGA	CGTCTCAATCTCGTATGCCGTCTTCTGCTTG
Round2_185	CAAGCAGAAGACGGCATACGAGATTGCAGAGAGA	CTCTGCAATCTCGTATGCCGTCTTCTGCTTG
Round2_186	CAAGCAGAAGACGGCATACGAGATAGGCACAAGA	TGTGCCTATCTCGTATGCCGTCTTCTGCTTG
Round2_187	CAAGCAGAAGACGGCATACGAGATAGTGTCCAGA	GGACACTATCTCGTATGCCGTCTTCTGCTTG
Round2_188	CAAGCAGAAGACGGCATACGAGATCAACCGTAGA	ACGGTTGATCTCGTATGCCGTCTTCTGCTTG
Round2_189	CAAGCAGAAGACGGCATACGAGATGTGTGCTAGA	AGCACACATCTCGTATGCCGTCTTCTGCTTG
Round2_190	CAAGCAGAAGACGGCATACGAGATCAGGACTAGA	AGTCCTGATCTCGTATGCCGTCTTCTGCTTG
Round2_191	CAAGCAGAAGACGGCATACGAGATACTTCCAAGA	TGGAAGTATCTCGTATGCCGTCTTCTGCTTG
Round2_192	CAAGCAGAAGACGGCATACGAGATATAGCAGAGA	CTGCTATATCTCGTATGCCGTCTTCTGCTTG

Description

#Round 2 barcode oligos

#Each barcode consists of one bottom strand oligo + it's matching top strand oligo which together will form the duplex

(rev. comp. part for most of the sequence, leaving some single stranded overhangs used for ligation)

#Round 2 barcodes are positioned in bases 11-17 of index 1, the actual barcode are the first seven bases of the oligo sequence

#Bases 8-10 in Index1 consist of the linker sequence 'TCT'

i7-Tru-Seq-long primer	Index PCR Reverse Primer	CAAGCAGAAGACGGCATACGAGAT

Barcoded Nextera Sequencing Primers for Index PCR (384 Primers used to barcode sub-libraries)		
Name	**Protocol Step**	**Sequence**
VS_Nextera_i5101	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACACGGTTCGTCGGCAGCGTC
VS_Nextera_i5102	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACAGGCATCGTCGGCAGCGTC
VS_Nextera_i5103	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACATCGCTCGTCGGCAGCGTC
VS_Nextera_i5104	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACCATCGTCGTCGGCAGCGTC
VS_Nextera_i5105	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACCGAACTCGTCGGCAGCGTC
VS_Nextera_i5106	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACGACTCTCGTCGGCAGCGTC
VS_Nextera_i5107	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACGCATGTCGTCGGCAGCGTC
VS_Nextera_i5108	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACGCCATTCGTCGGCAGCGTC
VS_Nextera_i5109	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACGCTGATCGTCGGCAGCGTC
VS_Nextera_i5110	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACGGAGTTCGTCGGCAGCGTC
VS_Nextera_i5111	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACTCAGCTCGTCGGCAGCGTC
VS_Nextera_i5112	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAACTTGCGTCGTCGGCAGCGTC
VS_Nextera_i5113	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGAAGCGTCGTCGGCAGCGTC
VS_Nextera_i5114	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGCCAGTTCGTCGGCAGCGTC
VS_Nextera_i5115	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGCCTTCTCGTCGGCAGCGTC
VS_Nextera_i5116	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGCTCAGTCGTCGGCAGCGTC
VS_Nextera_i5117	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGCTGCATCGTCGGCAGCGTC

VS_Nextera_i5118	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGGAACCTCGTCGGCAGCGTC
VS_Nextera_i5119	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGGCTAGTCGTCGGCAGCGTC
VS_Nextera_i5120	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGGTCGATCGTCGGCAGCGTC
VS_Nextera_i5121	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGTACGCTCGTCGGCAGCGTC
VS_Nextera_i5122	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGTCACGTCGTCGGCAGCGTC
VS_Nextera_i5123	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGTCGGATCGTCGGCAGCGTC
VS_Nextera_i5124	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAAGTGCCATCGTCGGCAGCGTC
VS_Nextera_i5125	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAATCCGCTTCGTCGGCAGCGTC
VS_Nextera_i5126	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAATCGTGGTCGTCGGCAGCGTC
VS_Nextera_i5127	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAATGACGGTCGTCGGCAGCGTC
VS_Nextera_i5128	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAATGCGTCTCGTCGGCAGCGTC
VS_Nextera_i5129	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAATGGCCTTCGTCGGCAGCGTC
VS_Nextera_i5130	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACAACTGGTCGTCGGCAGCGTC
VS_Nextera_i5131	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACAAGCGTTCGTCGGCAGCGTC
VS_Nextera_i5132	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACAAGGACTCGTCGGCAGCGTC
VS_Nextera_i5133	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACAATGCGTCGTCGGCAGCGTC
VS_Nextera_i5134	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACACAGGTTCGTCGGCAGCGTC
VS_Nextera_i5135	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACACTAGCTCGTCGGCAGCGTC
VS_Nextera_i5136	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACACTCCTTCGTCGGCAGCGTC
VS_Nextera_i5137	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACAGACACTCGTCGGCAGCGTC
VS_Nextera_i5138	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACAGCACATCGTCGGCAGCGTC
VS_Nextera_i5139	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACAGCCTTTCGTCGGCAGCGTC
VS_Nextera_i5140	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACAGGATGTCGTCGGCAGCGTC
VS_Nextera_i5141	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACATGCTCTCGTCGGCAGCGTC
VS_Nextera_i5142	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCAACCTTCGTCGGCAGCGTC
VS_Nextera_i5143	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCAATGCTCGTCGGCAGCGTC
VS_Nextera_i5144	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCACAAGTCGTCGGCAGCGTC
VS_Nextera_i5145	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCAGCAATCGTCGGCAGCGTC
VS_Nextera_i5146	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCAGGTTTCGTCGGCAGCGTC
VS_Nextera_i5147	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCATGGATCGTCGGCAGCGTC
VS_Nextera_i5148	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCGAAGATCGTCGGCAGCGTC
VS_Nextera_i5149	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCGATAGTCGTCGGCAGCGTC
VS_Nextera_i5150	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCTAGCATCGTCGGCAGCGTC
VS_Nextera_i5151	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCTGACTTCGTCGGCAGCGTC
VS_Nextera_i5152	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCTTCGTTCGTCGGCAGCGTC
VS_Nextera_i5153	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACCTTGACTCGTCGGCAGCGTC
VS_Nextera_i5154	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACGCAATCTCGTCGGCAGCGTC
VS_Nextera_i5155	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACGGAGAATCGTCGGCAGCGTC
VS_Nextera_i5156	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACGGTAGTTCGTCGGCAGCGTC
VS_Nextera_i5157	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACGTATGGTCGTCGGCAGCGTC
VS_Nextera_i5158	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACGTGAACTCGTCGGCAGCGTC
VS_Nextera_i5159	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACGTTGTGTCGTCGGCAGCGTC
VS_Nextera_i5160	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTACCACTCGTCGGCAGCGTC
VS_Nextera_i5161	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTCACCATCGTCGGCAGCGTC
VS_Nextera_i5162	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTCAGACTCGTCGGCAGCGTC
VS_Nextera_i5163	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTCCAGATCGTCGGCAGCGTC
VS_Nextera_i5164	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTCCTTGTCGTCGGCAGCGTC
VS_Nextera_i5165	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTCGAAGTCGTCGGCAGCGTC
VS_Nextera_i5166	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTCGGTATCGTCGGCAGCGTC
VS_Nextera_i5167	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTGAACGTCGTCGGCAGCGTC
VS_Nextera_i5168	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTGCGATTCGTCGGCAGCGTC
VS_Nextera_i5169	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTGGTGTTCGTCGGCAGCGTC
VS_Nextera_i5170	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTGTCAGTCGTCGGCAGCGTC
VS_Nextera_i5171	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTGTGCATCGTCGGCAGCGTC
VS_Nextera_i5172	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTTCACCTCGTCGGCAGCGTC
VS_Nextera_i5173	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACACTTGTCGTCGTCGGCAGCGTC
VS_Nextera_i5174	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGAACGGATCGTCGGCAGCGTC
VS_Nextera_i5175	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGAAGGTGTCGTCGGCAGCGTC
VS_Nextera_i5176	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGAAGTCCTCGTCGGCAGCGTC
VS_Nextera_i5177	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGAATCGGTCGTCGGCAGCGTC
VS_Nextera_i5178	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGACACAGTCGTCGGCAGCGTC
VS_Nextera_i5179	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGACATGCTCGTCGGCAGCGTC
VS_Nextera_i5180	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGAGACCATCGTCGGCAGCGTC
VS_Nextera_i5181	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGAGCTACTCGTCGGCAGCGTC
VS_Nextera_i5182	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGAGTACGTCGTCGGCAGCGTC
VS_Nextera_i5183	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGATCCTGTCGTCGGCAGCGTC
VS_Nextera_i5184	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGATCGCTTCGTCGGCAGCGTC
VS_Nextera_i5185	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGATGAGGTCGTCGGCAGCGTC
VS_Nextera_i5186	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCACACATCGTCGGCAGCGTC
VS_Nextera_i5187	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCATGAGTCGTCGGCAGCGTC
VS_Nextera_i5188	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCCAATGTCGTCGGCAGCGTC
VS_Nextera_i5189	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCCAGAATCGTCGGCAGCGTC
VS_Nextera_i5190	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCCTCTTTCGTCGGCAGCGTC
VS_Nextera_i5191	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCGAACTTCGTCGGCAGCGTC
VS_Nextera_i5192	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCGGTATTCGTCGGCAGCGTC
VS_Nextera_i5193	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCGTAACTCGTCGGCAGCGTC
VS_Nextera_i5194	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCGTGTATCGTCGGCAGCGTC
VS_Nextera_i5195	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCTACACTCGTCGGCAGCGTC
VS_Nextera_i5196	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCTCTAGTCGTCGGCAGCGTC
VS_Nextera_i5197	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGCTGTTCTCGTCGGCAGCGTC
VS_Nextera_i5198	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGAACTGTCGTCGGCAGCGTC
VS_Nextera_i5199	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGACATCTCGTCGGCAGCGTC
VS_Nextera_i5200	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGACCAATCGTCGGCAGCGTC
VS_Nextera_i5201	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGAGAAGTCGTCGGCAGCGTC
VS_Nextera_i5202	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGAGTGATCGTCGGCAGCGTC
VS_Nextera_i5203	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGATGGTTCGTCGGCAGCGTC

VS_Nextera_i5204	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGCAAGATCGTCGGCAGCGTC
VS_Nextera_i5205	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGCAGTTTCGTCGGCAGCGTC
VS_Nextera_i5206	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGCTACTTCGTCGGCAGCGTC
VS_Nextera_i5207	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGCTTACTCGTCGGCAGCGTC
VS_Nextera_i5208	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGTAGAGTCGTCGGCAGCGTC
VS_Nextera_i5209	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGTCTCATCGTCGGCAGCGTC
VS_Nextera_i5210	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGTGGTATCGTCGGCAGCGTC
VS_Nextera_i5211	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGGTTAGCTCGTCGGCAGCGTC
VS_Nextera_i5212	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTAGAGCTCGTCGGCAGCGTC
VS_Nextera_i5213	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTCACTCTCGTCGGCAGCGTC
VS_Nextera_i5214	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTCCTGTTCGTCGGCAGCGTC
VS_Nextera_i5215	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTCGGATTCGTCGGCAGCGTC
VS_Nextera_i5216	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTGAGGATCGTCGGCAGCGTC
VS_Nextera_i5217	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTGCAAGTCGTCGGCAGCGTC
VS_Nextera_i5218	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTGGCAATCGTCGGCAGCGTC
VS_Nextera_i5219	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTGGTTGTCGTCGGCAGCGTC
VS_Nextera_i5220	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTTACCGTCGTCGGCAGCGTC
VS_Nextera_i5221	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTTCGACTCGTCGGCAGCGTC
VS_Nextera_i5222	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACAGTTGCGTTCGTCGGCAGCGTC
VS_Nextera_i5223	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATACCAGGTCGTCGGCAGCGTC
VS_Nextera_i5224	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATACCGCATCGTCGGCAGCGTC
VS_Nextera_i5225	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATACGGTCTCGTCGGCAGCGTC
VS_Nextera_i5226	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATACGTCGTCGTCGGCAGCGTC
VS_Nextera_i5227	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATAGGCGATCGTCGGCAGCGTC
VS_Nextera_i5228	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATAGTGCCTCGTCGGCAGCGTC
VS_Nextera_i5229	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCAGACCTCGTCGGCAGCGTC
VS_Nextera_i5230	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCAGTGGTCGTCGGCAGCGTC
VS_Nextera_i5231	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCCAAGCTCGTCGGCAGCGTC
VS_Nextera_i5232	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCCTCCATCGTCGGCAGCGTC
VS_Nextera_i5233	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCCTGTGTCGTCGGCAGCGTC
VS_Nextera_i5234	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCGAGGTTCGTCGGCAGCGTC
VS_Nextera_i5235	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCGGAAGTCGTCGGCAGCGTC
VS_Nextera_i5236	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCGGCTTTCGTCGGCAGCGTC
VS_Nextera_i5237	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCGGTCATCGTCGGCAGCGTC
VS_Nextera_i5238	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATCTACGGTCGTCGGCAGCGTC
VS_Nextera_i5239	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGAAGGCTCGTCGGCAGCGTC
VS_Nextera_i5240	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGACTCGTCGTCGGCAGCGTC
VS_Nextera_i5241	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGAGCCTTCGTCGGCAGCGTC
VS_Nextera_i5242	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGCATCCTCGTCGGCAGCGTC
VS_Nextera_i5243	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGCCAACTCGTCGGCAGCGTC
VS_Nextera_i5244	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGCGATGTCGTCGGCAGCGTC
VS_Nextera_i5245	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGCGGAATCGTCGGCAGCGTC
VS_Nextera_i5246	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGCTCGTTCGTCGGCAGCGTC
VS_Nextera_i5247	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGGAAGGTCGTCGGCAGCGTC
VS_Nextera_i5248	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGGCACTTCGTCGGCAGCGTC
VS_Nextera_i5249	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGGTCACTCGTCGGCAGCGTC
VS_Nextera_i5250	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGTCCAGTCGTCGGCAGCGTC
VS_Nextera_i5251	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATGTGTGCTCGTCGGCAGCGTC
VS_Nextera_i5252	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATTACCGGTCGTCGGCAGCGTC
VS_Nextera_i5253	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATTAGGCGTCGTCGGCAGCGTC
VS_Nextera_i5254	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATTCCGAGTCGTCGGCAGCGTC
VS_Nextera_i5255	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATTCGCACTCGTCGGCAGCGTC
VS_Nextera_i5256	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACATTGCTGCTCGTCGGCAGCGTC
VS_Nextera_i5257	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAACACACTCGTCGGCAGCGTC
VS_Nextera_i5258	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAACCACATCGTCGGCAGCGTC
VS_Nextera_i5259	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAACCGATTCGTCGGCAGCGTC
VS_Nextera_i5260	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAAGAACGTCGTCGGCAGCGTC
VS_Nextera_i5261	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAAGCGTATCGTCGGCAGCGTC
VS_Nextera_i5262	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAAGGAACTCGTCGGCAGCGTC
VS_Nextera_i5263	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAAGTTGGTCGTCGGCAGCGTC
VS_Nextera_i5264	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAATACGGTCGTCGGCAGCGTC
VS_Nextera_i5265	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAATGCCTTCGTCGGCAGCGTC
VS_Nextera_i5266	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAATGGAGTCGTCGGCAGCGTC
VS_Nextera_i5267	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAATTGCCTCGTCGGCAGCGTC
VS_Nextera_i5268	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCACAATGGTCGTCGGCAGCGTC
VS_Nextera_i5269	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCACACCAATCGTCGGCAGCGTC
VS_Nextera_i5270	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCACAGAAGTCGTCGGCAGCGTC
VS_Nextera_i5271	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCACCAAGATCGTCGGCAGCGTC
VS_Nextera_i5272	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCACCATTCTCGTCGGCAGCGTC
VS_Nextera_i5273	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCACCTCATTCGTCGGCAGCGTC
VS_Nextera_i5274	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCACGAGAATCGTCGGCAGCGTC
VS_Nextera_i5275	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCACTCTCATCGTCGGCAGCGTC
VS_Nextera_i5276	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCACTGGTTTCGTCGGCAGCGTC
VS_Nextera_i5277	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGAAGGTTCGTCGGCAGCGTC
VS_Nextera_i5278	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGAGAGATCGTCGGCAGCGTC
VS_Nextera_i5279	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGAGTTGTCGTCGGCAGCGTC
VS_Nextera_i5280	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGATGAGTCGTCGGCAGCGTC
VS_Nextera_i5281	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGATTCCTCGTCGGCAGCGTC
VS_Nextera_i5282	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGCCTAATCGTCGGCAGCGTC
VS_Nextera_i5283	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGCTAACTCGTCGGCAGCGTC
VS_Nextera_i5284	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGGACTATCGTCGGCAGCGTC
VS_Nextera_i5285	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGGTACATCGTCGGCAGCGTC
VS_Nextera_i5286	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGGTGTTTCGTCGGCAGCGTC
VS_Nextera_i5287	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGTAGCATCGTCGGCAGCGTC
VS_Nextera_i5288	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGTCATCTCGTCGGCAGCGTC
VS_Nextera_i5289	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGTCCATTCGTCGGCAGCGTC

VS_Nextera_i5290	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGTGTGTTCGTCGGCAGCGTC
VS_Nextera_i5291	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCAGTTAGGTCGTCGGCAGCGTC
VS_Nextera_i5292	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATACACCTCGTCGGCAGCGTC
VS_Nextera_i5293	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATATCGGTCGTCGGCAGCGTC
VS_Nextera_i5294	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATCACCTTCGTCGGCAGCGTC
VS_Nextera_i5295	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATCCAAGTCGTCGGCAGCGTC
VS_Nextera_i5296	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATCGATCTCGTCGGCAGCGTC
VS_Nextera_i5297	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATCGGAATCGTCGGCAGCGTC
VS_Nextera_i5298	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATGAGTGTCGTCGGCAGCGTC
VS_Nextera_i5299	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATGATGCTCGTCGGCAGCGTC
VS_Nextera_i5300	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATGCAGATCGTCGGCAGCGTC
VS_Nextera_i5301	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATGGTCATCGTCGGCAGCGTC
VS_Nextera_i5302	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATTCTGGTCGTCGGCAGCGTC
VS_Nextera_i5303	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATTGACGTCGTCGGCAGCGTC
VS_Nextera_i5304	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCATTGCACTCGTCGGCAGCGTC
VS_Nextera_i5305	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCAAGGTATCGTCGGCAGCGTC
VS_Nextera_i5306	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCAATACCTCGTCGGCAGCGTC
VS_Nextera_i5307	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCACAATGTCGTCGGCAGCGTC
VS_Nextera_i5308	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCACAGAATCGTCGGCAGCGTC
VS_Nextera_i5309	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCACGAATTCGTCGGCAGCGTC
VS_Nextera_i5310	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCAGGTAATCGTCGGCAGCGTC
VS_Nextera_i5311	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCAGTAAGTCGTCGGCAGCGTC
VS_Nextera_i5312	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCATAGCTTCGTCGGCAGCGTC
VS_Nextera_i5313	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCATATGCTCGTCGGCAGCGTC
VS_Nextera_i5314	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCATCAACTCGTCGGCAGCGTC
VS_Nextera_i5315	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCATGACATCGTCGGCAGCGTC
VS_Nextera_i5316	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCATTGGATCGTCGGCAGCGTC
VS_Nextera_i5317	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCGAATACTCGTCGGCAGCGTC
VS_Nextera_i5318	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCGACAATTCGTCGGCAGCGTC
VS_Nextera_i5319	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCGATCAATCGTCGGCAGCGTC
VS_Nextera_i5320	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCGATTGTTCGTCGGCAGCGTC
VS_Nextera_i5321	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCGCATTATCGTCGGCAGCGTC
VS_Nextera_i5322	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCGGAATTTCGTCGGCAGCGTC
VS_Nextera_i5323	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCGTAAGATCGTCGGCAGCGTC
VS_Nextera_i5324	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCGTTGATTCGTCGGCAGCGTC
VS_Nextera_i5325	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCTACATGTCGTCGGCAGCGTC
VS_Nextera_i5326	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCTACGAATCGTCGGCAGCGTC
VS_Nextera_i5327	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCTAGAACTCGTCGGCAGCGTC
VS_Nextera_i5328	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCTAGTGATCGTCGGCAGCGTC
VS_Nextera_i5329	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCTCAAGTTCGTCGGCAGCGTC
VS_Nextera_i5330	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCTCTTAGTCGTCGGCAGCGTC
VS_Nextera_i5331	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCTGACAATCGTCGGCAGCGTC
VS_Nextera_i5332	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCTTAGAGTCGTCGGCAGCGTC
VS_Nextera_i5333	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCCTTCCTATCGTCGGCAGCGTC
VS_Nextera_i5334	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGAACAAGTCGTCGGCAGCGTC
VS_Nextera_i5335	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGAAGATCTCGTCGGCAGCGTC
VS_Nextera_i5336	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGAAGCAATCGTCGGCAGCGTC
VS_Nextera_i5337	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGACACTTTCGTCGGCAGCGTC
VS_Nextera_i5338	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGACGTTATCGTCGGCAGCGTC
VS_Nextera_i5339	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGACTAGTTCGTCGGCAGCGTC
VS_Nextera_i5340	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGAGAAGATCGTCGGCAGCGTC
VS_Nextera_i5341	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGAGATAGTCGTCGGCAGCGTC
VS_Nextera_i5342	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGAGTGATTCGTCGGCAGCGTC
VS_Nextera_i5343	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGATAACCTCGTCGGCAGCGTC
VS_Nextera_i5344	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGATAGTGTCGTCGGCAGCGTC
VS_Nextera_i5345	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGATCGAATCGTCGGCAGCGTC
VS_Nextera_i5346	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGATTCAGTCGTCGGCAGCGTC
VS_Nextera_i5347	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGCAACATTCGTCGGCAGCGTC
VS_Nextera_i5348	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGCAATCATCGTCGGCAGCGTC
VS_Nextera_i5349	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGCTACTATCGTCGGCAGCGTC
VS_Nextera_i5350	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGCTCAATTCGTCGGCAGCGTC
VS_Nextera_i5351	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGCTGTAATCGTCGGCAGCGTC
VS_Nextera_i5352	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGGAAGAATCGTCGGCAGCGTC
VS_Nextera_i5353	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGGATATGTCGTCGGCAGCGTC
VS_Nextera_i5354	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGGTATTCTCGTCGGCAGCGTC
VS_Nextera_i5355	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTACAGTTCGTCGGCAGCGTC
VS_Nextera_i5356	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTACTTCTCGTCGGCAGCGTC
VS_Nextera_i5357	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTAGACATCGTCGGCAGCGTC
VS_Nextera_i5358	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTAGCTTTCGTCGGCAGCGTC
VS_Nextera_i5359	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTAGTAGTCGTCGGCAGCGTC
VS_Nextera_i5360	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTCAGTATCGTCGGCAGCGTC
VS_Nextera_i5361	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTCATACTCGTCGGCAGCGTC
VS_Nextera_i5362	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTGCTAATCGTCGGCAGCGTC
VS_Nextera_i5363	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTGGAATTCGTCGGCAGCGTC
VS_Nextera_i5364	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTGTATCTCGTCGGCAGCGTC
VS_Nextera_i5365	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCGTGTTGTTCGTCGGCAGCGTC
VS_Nextera_i5366	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTAACGACTCGTCGGCAGCGTC
VS_Nextera_i5367	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTAAGTGCTCGTCGGCAGCGTC
VS_Nextera_i5368	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTAATCCGTCGTCGGCAGCGTC
VS_Nextera_i5369	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTACATGGTCGTCGGCAGCGTC
VS_Nextera_i5370	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTACTGCTTCGTCGGCAGCGTC
VS_Nextera_i5371	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTAGACTCTCGTCGGCAGCGTC
VS_Nextera_i5372	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTAGAGCATCGTCGGCAGCGTC
VS_Nextera_i5373	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTAGCTGTTCGTCGGCAGCGTC
VS_Nextera_i5374	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTATCTCGTCGTCGGCAGCGTC
VS_Nextera_i5375	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTATTCGCTCGTCGGCAGCGTC

VS_Nextera_i5376	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTCAAGTGTCGTCGGCAGCGTC
VS_Nextera_i5377	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTCACATCTCGTCGGCAGCGTC
VS_Nextera_i5378	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTCACTAGTCGTCGGCAGCGTC
VS_Nextera_i5379	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTCAGGAATCGTCGGCAGCGTC
VS_Nextera_i5380	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTCAGTCTTCGTCGGCAGCGTC
VS_Nextera_i5381	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTCTAACGTCGTCGGCAGCGTC
VS_Nextera_i5382	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTCTCAGATCGTCGGCAGCGTC
VS_Nextera_i5383	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTCTGCATTCGTCGGCAGCGTC
VS_Nextera_i5384	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTGACTGATCGTCGGCAGCGTC
VS_Nextera_i5385	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTGAGGTTTCGTCGGCAGCGTC
VS_Nextera_i5386	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTGATAGCTCGTCGGCAGCGTC
VS_Nextera_i5387	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTGCAACATCGTCGGCAGCGTC
VS_Nextera_i5388	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTGTACCTTCGTCGGCAGCGTC
VS_Nextera_i5389	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTGTCTACTCGTCGGCAGCGTC
VS_Nextera_i5390	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTGTGAAGTCGTCGGCAGCGTC
VS_Nextera_i5391	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTGTTCTGTCGTCGGCAGCGTC
VS_Nextera_i5392	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTTAGAGGTCGTCGGCAGCGTC
VS_Nextera_i5393	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTTCCACTTCGTCGGCAGCGTC
VS_Nextera_i5394	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTTCGCTATCGTCGGCAGCGTC
VS_Nextera_i5395	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTTCGTGTTCGTCGGCAGCGTC
VS_Nextera_i5396	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTTCTGGATCGTCGGCAGCGTC
VS_Nextera_i5397	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTTGACGTTCGTCGGCAGCGTC
VS_Nextera_i5398	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTTGATCGTCGTCGGCAGCGTC
VS_Nextera_i5399	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTTGCGTTTCGTCGGCAGCGTC
VS_Nextera_i5400	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACCTTGGTTCTCGTCGGCAGCGTC
VS_Nextera_i5401	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAACCAACTCGTCGGCAGCGTC
VS_Nextera_i5402	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAACGGTATCGTCGGCAGCGTC
VS_Nextera_i5403	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAACGTGTTCGTCGGCAGCGTC
VS_Nextera_i5404	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAACTTCGTCGTCGGCAGCGTC
VS_Nextera_i5405	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAAGACAGTCGTCGGCAGCGTC
VS_Nextera_i5406	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAAGCTCTTCGTCGGCAGCGTC
VS_Nextera_i5407	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAAGGACATCGTCGGCAGCGTC
VS_Nextera_i5408	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAAGGCTTTCGTCGGCAGCGTC
VS_Nextera_i5409	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAAGTGGATCGTCGGCAGCGTC
VS_Nextera_i5410	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAATAGGCTCGTCGGCAGCGTC
VS_Nextera_i5411	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAATCAGGTCGTCGGCAGCGTC
VS_Nextera_i5412	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAATGCGATCGTCGGCAGCGTC
VS_Nextera_i5413	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAATGTCCTCGTCGGCAGCGTC
VS_Nextera_i5414	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACACAGATCGTCGGCAGCGTC
VS_Nextera_i5415	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACAGGATTCGTCGGCAGCGTC
VS_Nextera_i5416	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACATACGTCGTCGGCAGCGTC
VS_Nextera_i5417	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACCTTGATCGTCGGCAGCGTC
VS_Nextera_i5418	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACGATGTTCGTCGGCAGCGTC
VS_Nextera_i5419	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACGGTAATCGTCGGCAGCGTC
VS_Nextera_i5420	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACTAGAGTCGTCGGCAGCGTC
VS_Nextera_i5421	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACTCACTTCGTCGGCAGCGTC
VS_Nextera_i5422	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACTCTACTCGTCGGCAGCGTC
VS_Nextera_i5423	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGACTGATCTCGTCGGCAGCGTC
VS_Nextera_i5424	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGACAAGTCGTCGGCAGCGTC
VS_Nextera_i5425	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGAGCAATCGTCGGCAGCGTC
VS_Nextera_i5426	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGATTGGTCGTCGGCAGCGTC
VS_Nextera_i5427	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGCATCTTCGTCGGCAGCGTC
VS_Nextera_i5428	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGCCATATCGTCGGCAGCGTC
VS_Nextera_i5429	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGGAAGATCGTCGGCAGCGTC
VS_Nextera_i5430	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGGTATGTCGTCGGCAGCGTC
VS_Nextera_i5431	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGGTCATTCGTCGGCAGCGTC
VS_Nextera_i5432	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGTACTGTCGTCGGCAGCGTC
VS_Nextera_i5433	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGTCGTTTCGTCGGCAGCGTC
VS_Nextera_i5434	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGTGTAGTCGTCGGCAGCGTC
VS_Nextera_i5435	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGAGTTGACTCGTCGGCAGCGTC
VS_Nextera_i5436	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATACCGTTCGTCGGCAGCGTC
VS_Nextera_i5437	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATACGTGTCGTCGGCAGCGTC
VS_Nextera_i5438	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATAGTGCTCGTCGGCAGCGTC
VS_Nextera_i5439	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATCAACCTCGTCGGCAGCGTC
VS_Nextera_i5440	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATCAGGTTCGTCGGCAGCGTC
VS_Nextera_i5441	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATCCTCATCGTCGGCAGCGTC
VS_Nextera_i5442	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATGACCATCGTCGGCAGCGTC
VS_Nextera_i5443	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATGCGAATCGTCGGCAGCGTC
VS_Nextera_i5444	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATGGAAGTCGTCGGCAGCGTC
VS_Nextera_i5445	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATGTAGCTCGTCGGCAGCGTC
VS_Nextera_i5446	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATGTGCTTCGTCGGCAGCGTC
VS_Nextera_i5447	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATTCCAGTCGTCGGCAGCGTC
VS_Nextera_i5448	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGATTGGCATCGTCGGCAGCGTC
VS_Nextera_i5449	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCAACATCTCGTCGGCAGCGTC
VS_Nextera_i5450	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCAACCAATCGTCGGCAGCGTC
VS_Nextera_i5451	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCAAGAAGTCGTCGGCAGCGTC
VS_Nextera_i5452	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCAAGTCTTCGTCGGCAGCGTC
VS_Nextera_i5453	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCACAAGATCGTCGGCAGCGTC
VS_Nextera_i5454	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCACATAGTCGTCGGCAGCGTC
VS_Nextera_i5455	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCACTCTATCGTCGGCAGCGTC
VS_Nextera_i5456	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCACTGATTCGTCGGCAGCGTC
VS_Nextera_i5457	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCAGAGTTTCGTCGGCAGCGTC
VS_Nextera_i5458	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCAGCTTATCGTCGGCAGCGTC
VS_Nextera_i5459	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCAGTACTTCGTCGGCAGCGTC
VS_Nextera_i5460	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCAGTTACTCGTCGGCAGCGTC
VS_Nextera_i5461	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCATAACGTCGTCGGCAGCGTC

VS_Nextera_i5462	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCATGGAATCGTCGGCAGCGTC
VS_Nextera_i5463	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCATTGTCTCGTCGGCAGCGTC
VS_Nextera_i5464	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCCAAGAATCGTCGGCAGCGTC
VS_Nextera_i5465	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCCAGTTATCGTCGGCAGCGTC
VS_Nextera_i5466	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCCATAACTCGTCGGCAGCGTC
VS_Nextera_i5467	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCCTATTGTCGTCGGCAGCGTC
VS_Nextera_i5468	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCCTTACATCGTCGGCAGCGTC
VS_Nextera_i5469	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCGAACATTCGTCGGCAGCGTC
VS_Nextera_i5470	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCGAATCATCGTCGGCAGCGTC
VS_Nextera_i5471	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCGTAGTATCGTCGGCAGCGTC
VS_Nextera_i5472	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCGTCTAATCGTCGGCAGCGTC
VS_Nextera_i5473	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCGTGATTTCGTCGGCAGCGTC
VS_Nextera_i5474	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCGTTAAGTCGTCGGCAGCGTC
VS_Nextera_i5475	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTAACTCTCGTCGGCAGCGTC
VS_Nextera_i5476	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTAATGGTCGTCGGCAGCGTC
VS_Nextera_i5477	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTACACATCGTCGGCAGCGTC
VS_Nextera_i5478	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTAGAGTTCGTCGGCAGCGTC
VS_Nextera_i5479	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTATGAGTCGTCGGCAGCGTC
VS_Nextera_i5480	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTCCAATTCGTCGGCAGCGTC
VS_Nextera_i5481	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTCGTAATCGTCGGCAGCGTC
VS_Nextera_i5482	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTCTATGTCGTCGGCAGCGTC
VS_Nextera_i5483	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTCTTGTTCGTCGGCAGCGTC
VS_Nextera_i5484	Index PCR Forward Primer	AATGATACGGCGACCACCGAGATCTACACGCTGGATATCGTCGGCAGCGTC

Chapter 3: Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells

Status in submission process: Available on biorxiv; ready for submission

Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells

Volker Soltys[1*], Moritz Peters[1], Dingwen Su[1], Marek Kučka[1,2,], Yingguang Frank Chan[1,3]

1 Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany
2 Department of Translational Genomics, University of Cologne, 50931 Cologne, Germany
3 University of Groningen, Groningen Institute of Evolutionary Life Sciences, 9747 AG Groningen, Netherlands

* Corresponding authors
volker.soltys@tue.mpg.de; frank.chan@rug.nl

Abstract

Gene expression and chromatin accessibility are highly interconnected processes. Disentangling one without the other provides an incomplete picture of gene regulation. However, simultaneous measurements of RNA and accessible chromatin are technically challenging, especially when studying complex organs with rare cell-types. Here, we present easySHARE-seq, an elaboration of SHARE-seq, providing simultaneous measurements of ATAC- and RNA-seq within single cells, enabling identification of cell-type specific *cis*-regulatory elements (CREs). easySHARE-seq retains high scalability, improves RNA-seq data quality while also allowing for flexible study design. Using 19,664 joint profiles from murine liver nuclei, we linked CREs to their target genes and uncovered complex regulation of key genes such as *Gata4*. We further identify *de novo* genes and *cis*-regulatory elements displaying zonation in Liver sinusoidal epithelial cells (LSECs), a challenging cell type with low mRNA levels, demonstrating the power of multimodal measurements. EasySHARE-seq therefore provides a flexible platform for investigating gene regulation across cell types and scale.

Introduction

Gene expression and chromatin state together influence fundamental processes such as gene regulation or cell fate decisions [1–3]. A better understanding of these mechanisms and their interactions will be a major step toward decoding developmental trajectories or reconstructing cellular taxonomies in both health and disease. However, to fully capture these complex relationships, multiple information layers need to be measured simultaneously. For example, prior studies have argued that chromatin state is often predictive of gene expression and can also prime cells toward certain lineage decisions or even induce tissue regeneration[4–6]. However, these studies depend on the computational integration of separately measured modalities. By assuming a shared biological state, this restricts the discovery of novel and potentially fine-scale differences and renders it challenging to identify the root cause of erroneous cell states[7].

The last decade has seen an explosive growth in single-cell methodologies, with new assays, increasing throughput and a suite of computational tools[8]. Most non-commercial high-throughput methodologies rely on combinatorial indexing for single-cell barcoding, where sequential rounds of barcodes combine to create unique cellular barcode combinations [9,10]. Compared to single-modality assays, multi-omic technologies, which capture two or more information layers, are relatively new. Therefore, they are still limited in sensitivity and throughput and commercial kits can be expensive such that multi-omic studies tend to have limited sample sizes [11,12].

To address these problems, we built upon the previously published protocol called SHARE-seq[13] and developed easySHARE-seq, a protocol for simultaneously measuring gene expression and chromatin accessibility within single cells using combinatorial indexing. Major improvements include easySHARE-seq's barcoding framework, which allows for expanded and flexible study design, all while being compatible with standard Illumina sequencing, thereby removing economic hurdles. Importantly, easySHARE-seq retains the scalability and improves upon RNA-seq sensitivity of the original SHARE-seq protocol. Here, we used easySHARE-seq to profile 19,664 murine liver nuclei and show that we can recover high quality data in both RNA-seq and ATAC-seq channels, which are highly congruent and share equal power in classifying cell types. We then surveyed the *cis*-regulatory landscape of Liver Sinusoidal Endothelial Cells (LSECs), leveraging the simultaneous measurements of gene expression and chromatin accessibility and identified 40,957 links between expressed genes and nearby ATAC-seq peaks. Notably, genes with the highest number of links were enriched for transcription factors and regulators known to control important functions within LSECs. Lastly, we show that easySHARE-seq can be used to investigate micro-scale changes in accessibility and gene expression by identifying novel markers and open chromatin regions displaying zonation in LSECs. This technology improves our toolkit of multi-omic protocols needed for advancing our knowledge about gene regulation and cell fate decisions.

Results

easySHARE-seq reliably labels both transcriptome and accessible chromatin in individual cells

To develop a multi-omic single-cell (sc) RNA and scATAC-seq protocol that allows for flexible study design while being highly scalable, we built upon SHARE-seq[13] to create easySHARE-seq, which uses two rounds of ligation to simultaneously label cDNA and DNA fragments in the same cell (**Fig. 1A**). Due to a much more streamlined barcoding structure, easySHARE-seq allows 300bp sequencing of the insert. This longer read-length leads to a higher recovery of DNA variants, thus increasing the power to detect allele-specific signals or cell-specific variation, e.g., in hybrids or cancer cells[14].

To generate libraries, fixed and permeabilized cells or nuclei (we will use “cells” afterwards to refer to both) are transposed by Tn5 transposase carrying a custom adapter with a single-stranded overhang (**Fig. 1B**). Next, mRNA is reverse transcribed (RT) using a biotinylated poly(T) primer with an identical overhang. Subsequently, the cells are individually barcoded in two rounds of combinatorial indexing with 192 barcodes in each round, creating a total of 36,864 possible barcode combinations.The first barcode is ligated onto the already present overhang and itself contains a second single-stranded overhang, onto which the second barcode can be ligated. Importantly, in the easySHARE-seq design, we have kept the total length of the barcode within 17nt (“Index 1” read; **Fig. 1B, Suppl. Fig. 1A**), allowing for multiplexing of easySHARE-seq libraries with standard Illumina libraries. In contrast, in the original publication, SHARE-seq libraries required Index 1 lengths of 99nt, a highly custom configuration which would require a costly private sequencing.

After barcoding, the cells are aliquoted into sub-libraries of approximately 3,500 cells each and reverse crosslinked. A streptavidin pull-down of the biotinylated RT-primer is performed to separate the cDNA molecules from the chromatin (“fragments”). Each sub-library is then prepared for sequencing and amplified using matched indexing primers to allow identification of paired cellular scRNA- and scATAC-seq profiles. By scaling up the numbers of sub-libraries, this barcoding strategy therefore allows for high-throughput experiments of hundreds of thousands of cells, only limited by the availability of indexing primers. For a detailed description of the flexibility of easySHARE-seq, instructions on how to modify and incorporate the framework into new designs as well as critical steps to assess when planning to use easySHARE-seq see Supplementary Notes.

To evaluate the accuracy and cell-specificity of the barcoding, we first performed easySHARE-seq on a mixed pool between human and murine cell lines (HEK and OP-9 respecitvely). This design allows us to identify two or more cells sharing the same barcode (‘doublets’; **Fig. 1C**, left). After sequencing, we recovered a total of 3,808 cells. Both chromatin and transcriptome profiles separated well within each cell (**Fig. 1C**, middle), with cDNA showing a lower accuracy with increasing transcript counts, likely due to less precise read mapping. We identified a total of 124 doublets (**Fig, 1C**, right), which gives a final doublet rate of 6.34% factoring in the undetectable intra-species doublets. For comparison, a 10X Chromium Next GEM experiment with 10,000 cells has a doublet rate of ~7.9% (www.10xgenomics.com). Importantly, easySHARE-seq doublet rates can be lowered further by aliquoting fewer cells within each sub-library. To summarise, easySHARE-seq provides a high-throughput and flexibility framework for accurately measuring chromatin accessibility and gene expression in single cells.

Simultaneous scATAC-seq and scRNA-seq profiling in murine primary liver cells

To assess data quality and investigate the relationship between gene expression and chromatin accessibility, we focused on murine liver. The liver consists of a diverse set of defined primary cell types, ranging from large and potentially multinucleated hepatocytes to small non-parenchymal cell types such as Liver Sinusoidal Endothelial Cells[15] (LSECs).

We generated matched high-quality chromatin and gene expression profiles for 19,664 adult liver cells across four age-matched mice (2 male, 2 female), amounting to a recovery rate of 70.2% (28,000 input cells). Each nuclei had on average 3,629 UMIs and 2,213 fragments (74% of all RNA-seq reads were cDNA, 55.9% mean ATAC-seq fragments in peaks; **Suppl. Fig. 1B & D**). In terms of UMIs per cell, easySHARE-seq therefore out-performed other previously published multi-omic and representative single channel assays (**Fig. 2B**; see figure legend for tissue type and study). Consistent with nuclei as input material, the majority of cDNA molecules were intronic (69.6%, **Suppl. Fig. 1C & H**). Regarding DNA fragments per cell, easySHARE-seq performed similarly to other published multi-omic assays (**Fig. 2C**) and scATAC-seq libraries displayed the characteristic banding pattern with reads being highly enriched at transcription start sites (TSS; **Suppl. Fig. 1 E, F, H**).

To visualise and identify cell types, we first projected the ATAC- and RNA-seq modalities separately into 2D Space and then used Weighted Nearest Neighbor[16] (WNN) integration to combine both modalities into a single UMAP visualisation (**Fig. 2A)**. Importantly, the same cells independently clustered together in the scRNA- and scATAC-seq UMAPs, showcasing high congruence between the two modalities (**Suppl. Fig. 2A&B**). We then annotated previously published cell types based on gene expression of previously established marker genes [17,18]. Marker gene expression was highly specific to the clusters (**Fig. 2D, Suppl. Fig. 2F**) and we recovered all expected cell types (**Suppl. Fig.2C**). Importantly, the same cell types were identified using each modality independently, showcasing high congruence between the scATAC- and scRNA-seq modalities (**Fig. 2E**). Altogether, our results show that easySHARE-seq generates high quality joint cellular profiles of chromatin accessibility and gene expression within primary tissue, expanding our toolkit of multi-omic protocols.

Uncovering the cis-regulatory landscape of key regulators through peak-gene associations

As easySHARE-seq simultaneously measures chromatin accessibility and gene expression, it allows to direct investigation of the relationship between them to potentially connect *cis*-regulatory elements (CREs) to their target genes. To do so, we adopted the analytical framework from Ma et al.[13], which queries if an increased expression within a cell is significantly correlated with chromatin accessibility at a peak while controlling for GC content and accessibility strength. Focusing on LSECs (1,501 cells), we calculated associations between putative CREs (pCREs, defined as peaks with a significant peak–gene association) and each expressed gene, considering all peaks within ± 500kb of the TSS. We identified 40,957 significant peak–gene associations (45% of total peaks, $P < 0.05$, FDR = 0.1) with 15,061 genes having at least one association (76.8% of all expressed genes, **Suppl. Fig. 3A,C**). Conversely, some rare pCREs (2.9%) were associated with five or more genes (0.03% when considering only pCREs within ± 50 kb of a TSS (**Suppl. Fig. 3B,D**)). These pCREs tended to cluster to regions of higher expressed gene density (2.15 mean expressed genes within 50kbp vs 0.93 for all global peaks) and their associated genes were enriched for biological processes such as mRNA processing, histone modifications and splicing (**Suppl. Fig. 3H**), possibly reflecting loci with increased regulatory activity.

Focusing on genes, we ranked them based on their number of associated pCREs (**Fig. 2F**). Within the top 1% genes with the most pCRE associations were many key regulators and transcription factors. Examples include *Taf5,* which directly binds the TATA-box[19] and is required for initiation of transcription, or *Gata4,* which has been identified as the master regulator for LSEC specification during development as well as controlling regeneration and metabolic maturation of liver tissue in adult mice [20,21]. As such, it incorporates a variety of signals and its expression needs to be strictly regulated, which is reflected in its many pCREs associations (**Fig. 2H**). Similarly, *Igf1* also integrates signals from many different pCREs[22] (**Suppl. Fig. 3G**). Notably, pCRES are significantly enriched at transcription start sites (TSS), even relative to background enrichment (**Fig. 2G**).
To summarise, easySHARE-seq allows the direct investigation of the relationship between chromatin accessibility and gene expression and identify putative *cis*-regulatory elements at genomic scale, even in small cell types with relatively low mRNA contents (**Suppl. Fig. 2D)**.

***De novo* identification of open chromatin regions and genes displaying zonation in LSECs**

We next investigated the process of zonation in LSECs. The liver consists of hexagonal units called lobules where blood flows from the portal vein and arteries toward a central vein [23,24] (**Fig. 3A)**. The central–portal (CP) axis is characterised by a morphogen gradient, e.g. *Wnt2*, secreted by central vein LSECs, with the resulting micro-environment giving rise to spatial division of labour among hepatocytes [25–27]. Studying zonation in non-parenchymal cells such as LSECs is challenging as these are small cells with low mRNA content (**Suppl. Fig. 2D,E**), lying below the detection limit of current spatial transcriptomic techniques. As a result, only very few studies assess zonation in LSECs on a genomic level [28]. However, LSECs are critical to liver function as they line the artery walls, clear and process endotoxins, play a critical role in liver regeneration and secrete morphogens themselves to regulate hepatocyte gene expression [29–31], rendering their understanding a prerequisite for tackling many diseases.
We therefore asked if we can recover known zonation gradients and potentially identify novel marker genes and open chromatin regions displaying zonation. We noticed that LSECs clustered in a distinct linear pattern in our UMAP projection and therefore divided them into equal bins along UMAP2 coordinates (**Suppl. Fig. 4A**, number of cells per bin 80-260, median: 128). We then calculated mean normalised expression and mean normalised accessibility within each bin. This recovered gene expression and chromatin accessibility gradients for major known zonation marker genes[28] (**Fig. 3B,C**). For example, *Wnt2* expression decreased strongly along the CP axis as did chromatin accessibility of all three peaks at the *Wnt2* locus (**Fig. 3B**). We also recovered the zonation profiles for the majority of known pericentral (increasing along the CP-axis), periportal (decrease along the CP-axis) and non-monotonic markers (decrease toward both ends) as well as their associated chromatin regions (**Fig. 3C**). Gene expression zonation profiles can also be recovered by ordering LSECs along pseudotime (**Suppl. Fig. 4C,D**). In contrast, simply subclustering LSECs and comparing expression between these clusters was too broad for the assessment of zonation (**Suppl. Fig. 4A,B**).
Next, we sought to identify novel marker genes and open chromatin regions displaying zonation in LSECs based on the decrease or increase of mean expression or accessibility along the previously established bins. In total, we classified 153 genes and 381 open chromatin regions as pericentral and 209 genes and 465 open chromatin regions showed periportal zonation profiles (**Fig. 3D**). The list of markers contained many genes regulating epithelial growth and angiogenesis (e.g. *Efna1, Nrg2, Zfpm1, Zfpm2, Bmpr2*)[32–34], related to

regulating hepatocyte functions and communication (e.g. *Dll4, Foxo1, Sp1, Snx3*)[35–37] as well as immunological functions (e.g. *Sirt2, Cd59a)*[38,39]*,* suggesting that these processes show variation along the PC axis. As dysregulation of LSEC zonation is implicated in multiple illnesses such as liver cirrhosis or non-alcoholic fatty liver disease [40,41], these genes are potential new biomarkers for their identification and the open chromatin regions starting points for investigating the role of gene regulation in their emergence.

Discussion

Understanding complex processes such as gene regulation or disease states requires the integration of multiple layers of information. Here, we show that easySHARE-seq provides a high-quality, high-throughput and flexible platform for joint profiling of chromatin accessibility and gene expression within single cells. We show that both modalities are highly congruent with one another and we leverage their simultaneous measurements to identify peak–gene interactions and survey the *cis*-regulatory landscape of LSECs. We also show that easySHARE-seq can be used to assess micro-scale changes such as zonation in LSECs across both gene expression and chromatin accessibility. These cells have low mRNA content and we recovered zonation profiles of many transcription factors, which are often lowly expressed, further demonstrating the power of easySHARE-seq.

Besides improving upon RNA-seq data quality, we argue that easySHARE-seq has many advantages, especially in terms of the sequencing flexibility due to the barcode design, which can help remove hurdles for incorporating multi-omic single-cell assays into study designs. Combined with shorter experimental times (~12h total), easySHARE-seq might be particularly suited for studies where higher sample sizes are required or ones that rely on identification of genomic variants, e.g., in diverse, non-inbred individuals or in cancer. In terms of costs per cell, easySHARE-seq performs similarly to standard SHARE-seq with ~0.056 cents/cell, a fraction of the costs (<25%) of commercially available platforms, even before factoring in the specialized instrument costs. A comparison between technologies can be found in **Table 1**.

We envision easySHARE-seq as another technological step toward ultimately understanding gene regulation in health and disease, surveying *cis*-regulatory landscapes during differentiation and lineage commitment and determining genetic variants affecting those processes.

Figure 1

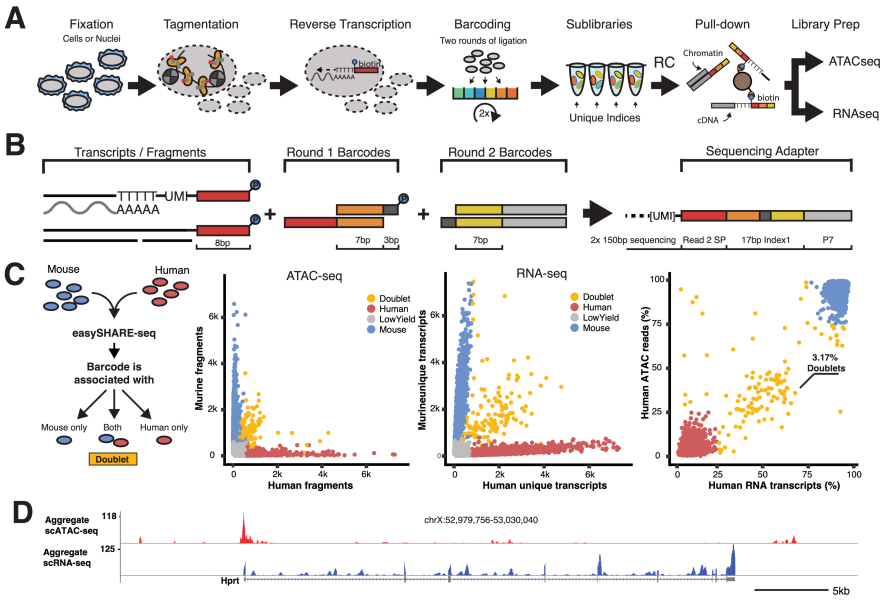


Figure 1: easySHARE-seq enables highly-accurate simultaneous scATAC-seq and scRNA-seq profiling

(A) Schematic workflow of easySHARE-seq.
(B) Generation and structure of the single-cell barcoding within Index 1.
(C) Principle of a species-mixing experiment. Cells are mixed prior to easySHARE-seq and sequences associated with each cell barcode are assessed for genome of origin (left panel). Unique ATAC fragments per cell aligning to the mouse or human genome (middle left). Cells are coloured according to their assigned origin (red: human; blue: mouse; orange: doublet). Middle right: Same plot but with RNA UMIs. Right: Percentage of ATAC fragments or RNA UMIs per cell relative to total sequencing reads mapping uniquely to the human genome. 3.17% of all observed cells classified as doublets. Accounting for same-species doublets, this results in a doublet rate of 6.34%.
(D) Aggregate chromatin accessibility (red) and expression-seq (blue) profile of OP-9 cells at the *Hprt* locus.

Figure 2

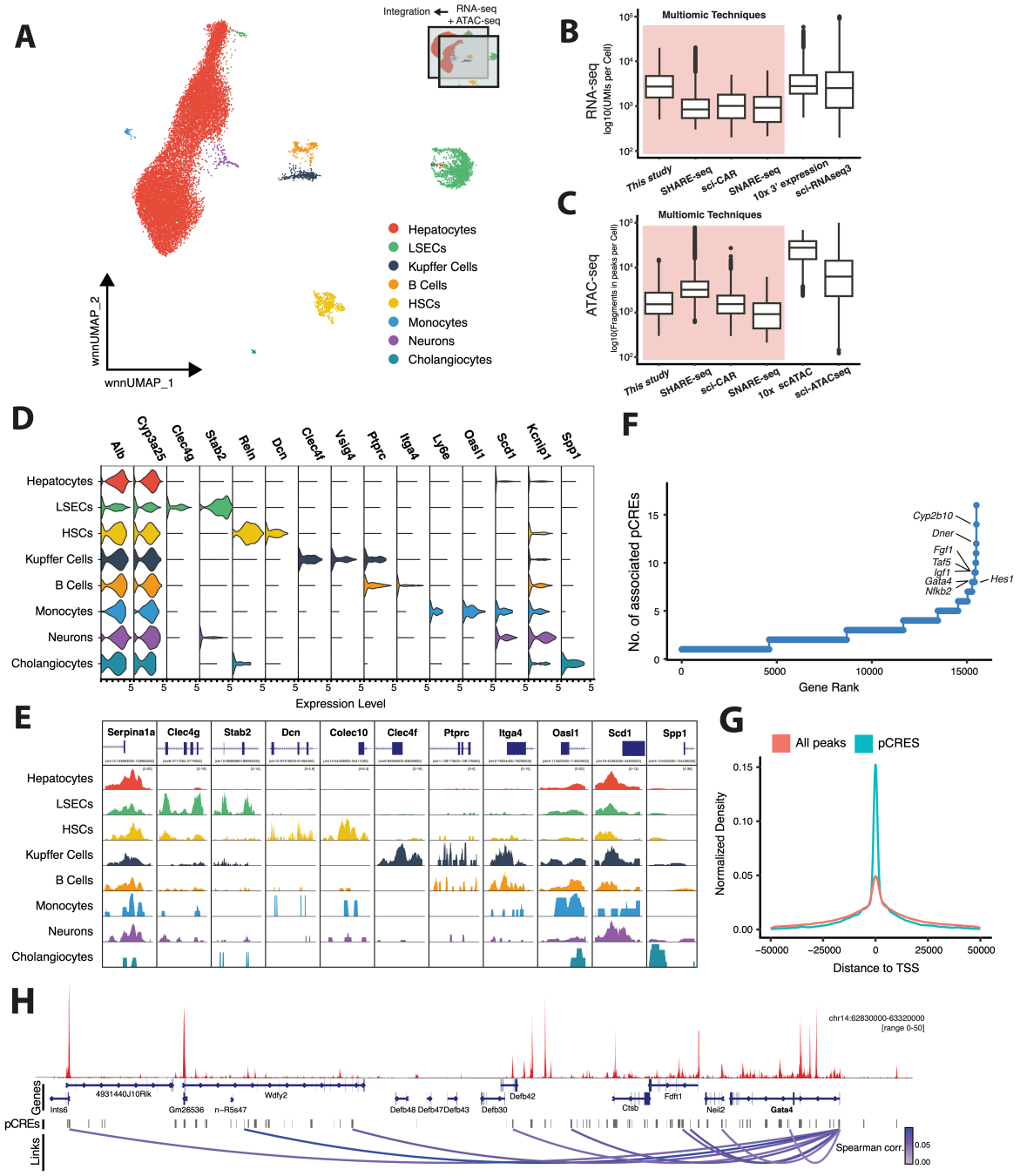

A
Integration
RNA-seq
+ ATAC-seq
Hepatocytes
LSECs
Kupffer Cells
B Cells
HSCs
Monocytes
Neurons
Cholangiocytes
wnnUMAP_2
wnnUMAP_1
B
Multiomic Techniques
RNA-seq
log10(UMIs per Cell)
This study
SHARE-seq
sci-CAR
SNARE-seq
10x 3' expression
sci-RNAseq3
C
Multiomic Techniques
ATAC-seq
log10(Fragments in peaks per Cell)
This study
SHARE-seq
sci-CAR
SNARE-seq
10x scATAC
sci-ATACseq
D
Alb
Cyp3a25
Clec4g
Stab2
Reln
Dcn
Clec4f
Vsig4
Ptprc
Itga4
Ly6e
Oasl1
Scd1
Kcnip1
Spp1
Hepatocytes
LSECs
HSCs
Kupffer Cells
B Cells
Monocytes
Neurons
Cholangiocytes
Expression Level
E
Serpina1a
Clec4g
Stab2
Dcn
Colec10
Clec4f
Ptprc
Itga4
Oasl1
Scd1
Spp1
Hepatocytes
LSECs
HSCs
Kupffer Cells
B Cells
Monocytes
Neurons
Cholangiocytes
F
No. of associated pCREs
Gene Rank
Cyp2b10
Dner
Fgf1
Tat5
Igf1
Gata4
Nfkb2
Hes1
G
All peaks
pCRES
Normalized Density
Distance to TSS
H
chr14:62830000-63320000
[range 0-50]
Genes
pCREs
Links
4931440J10Rik
Wdfy2
Ints6
Gm26536
n-R5s47
Defb48
Defb47
Defb43
Defb30
Defb42
Ctsb
Fdft1
Neil2
Gata4
Spearman corr.

Figure 2: Joint expression and chromatin accessibility profiling in primary liver nuclei

(A) UMAP visualisation of WNN-integrated scRNAseq and scATACseq modalities of 19,664 liver nuclei. Nuclei are coloured by cell types.
(B) Comparison of UMIs/cell across different single-cell technologies. Red shading denotes all multi-omic technologies. Datasets are this study, SHARE-seq[13] (murine skin cells), sci-CAR[11] (murine kidney nuclei), SNARE-seq[12] (adult & neonatal mouse cerebral cortex nuclei), 10x 3' Expression[17] (murine liver nuclei) and sci-RNAseq3 [9](E16.5 mouse embryo nuclei).
(C) Comparison of unique fragments per cell across different single-cell technologies. Colouring as in (B). Datasets differing to (B) are 10x 3'scATAC[42] (murine liver nuclei) and sciATAC-seq[43] (murine liver nuclei).
(D) Normalised gene expression of representative marker genes per cell type.
(E) Aggregate ATAC-seq tracks at marker accessibility peaks per cell type.
(F) Genes ranked by number of significantly correlated pCREs ($P < 0.05$, FDR = 0.1) per gene (±500kbp from TSS) in LSECs. Marked are transcription factors & regulators within the top 1% of genes with a critical role in LSECs.
(G) Significantly correlated pCREs are enriched for TSS proximity. Normalised density of all peaks versus pCREs within ±50kbp of nearest TSS.
(H) Aggregate scATAC-seq track of LSECs at the *Gata4* locus and 500kbp upstream region. Loops denote pCREs significantly correlated with *Gata4* and are coloured by Spearman correlation of respective pCRE–*Gata4* comparison

Figure 3

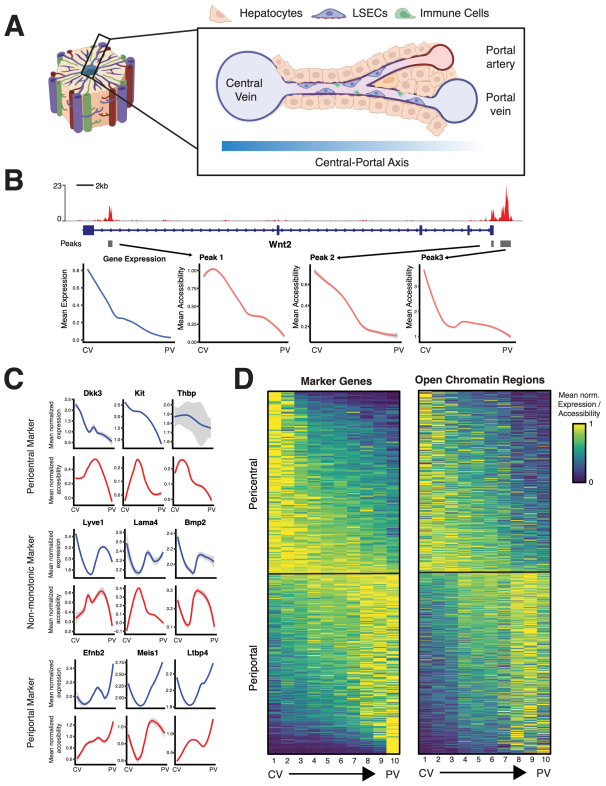


Figure 3: Zonation profiles in LSECs across gene expression and chromatin accessibility

(A) Schematic of a liver lobule. A liver lobule has a 'Central–Portal Axis' starting from the central vein to the portal vein and portal artery. The sinusoidal capillary channels are lined with LSECs.

(B) Changes along the Central–Portal Axis at the *Wnt2* locus. Top: Aggregate scATACseq profile (red) of LSECs at *Wnt2* locus. Grey bars denote identified peaks. Bottom: In blue, loess trend line of mean normalised *Wnt2* gene expression along the Central–Portal-Axis (central vein, CV; portal vein, PV; split into equal 10 bins). In red, loess trend line of mean normalised chromatin accessibility in peaks at the *Wnt2* locus along the CP-axis.

(C) Loess trend line of mean normalised expression (blue) and mean normalised accessibility along the Central–Portal axis for pericentral markers (top, increased toward the central vein, *Dkk3*, *Kit* and *Thbp*), non-monotonic markers (middle, increased between the veins, *Lyve1*, *Lama4* and *Bmp2*) and periportal markers (increased toward the portal vein, *Efnb2*, *Meis1* & *Ltbp4*)

(D) Left: Zonation profiles of 362 genes along the Central–Portal axis. Right: Zonation profiles of 846 open chromatin regions along the Central–Portal axis. All profiles are normalised by their maximum.

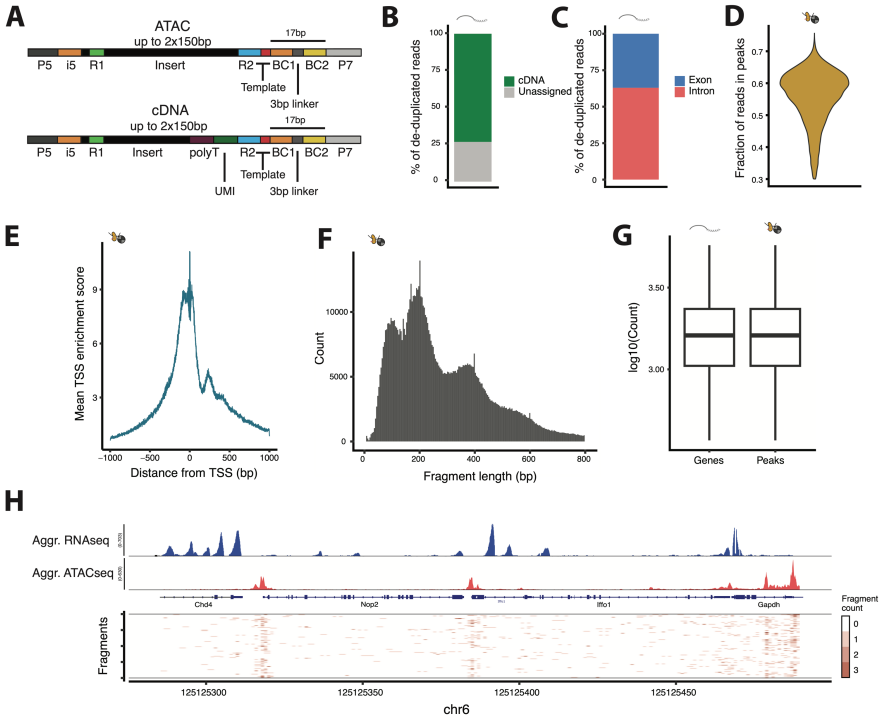
Table 1

Comparison of single-cell techniques

	Cost / Cell	Throughput	Multiomic?	Special equipment?	Std. sequencing?	Potential insert length?
This study	5.6 ct	> 200.000	Yes	No	Yes	> 200bp
SHARE-seq	4.33 ct	> 200.000	Yes	No	No	100bp
10x Multiome	25.8 ct	80.000	Yes	Yes	No	100bp
sci-RNA-seq3	1 ct	> 200.000	No	No	Yes	> 200bp

Table 1: Comparison between different single-cell technologies

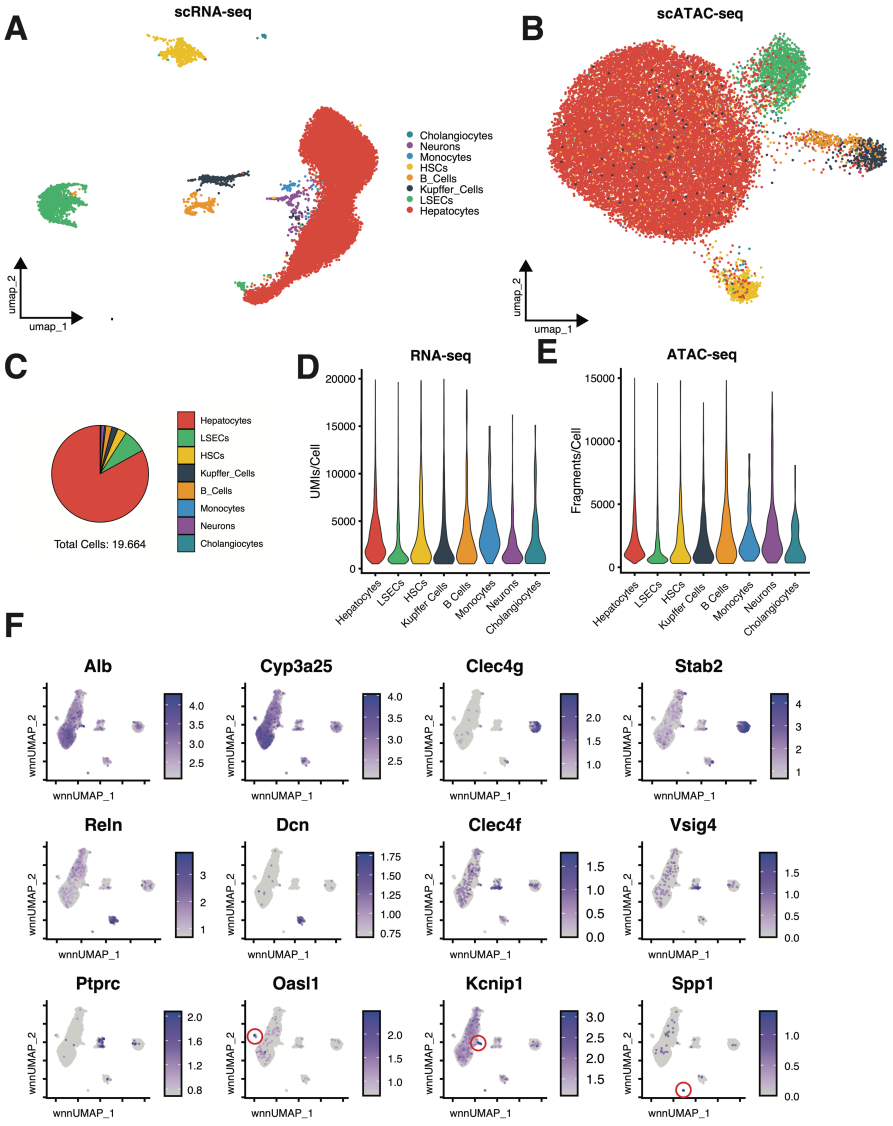
Supplementary Figure 1



Supplementary Figure 1: Barcode structure and summary of quality control measures in liver nuclei

(A) Structure of a scATAC-seq and scRNA-seq sequencing read. Created with Biorender.com
(B) Percentage of total scRNAseq sequencing reads containing cDNA fragments.
(C) Percentage of de-duplicated scRNAseq sequencing reads overlapping an exon or intron.
(D) Distribution of fraction of reads in peaks (FRiP) per cell in the scATAC-seq data (mean: 0.55).
(E) Mean TSS enrichment score per cell in relation to distance from nearest TSS in the scATACseq data.
(F) Histogram of fragment length in scATAC sequencing reads
(G) Expressed genes and accessible peaks per cell (mean expressed genes: 1,798; mean accessible peaks: 1,983)
(H) Top: Aggregate scRNA-seq (blue) and scATAC-seq (red) of all liver nuclei at *Nop2/Iffo2/Gapdh* locus. Bottom: Chromatin accessibility profiles of 100 individual cells.

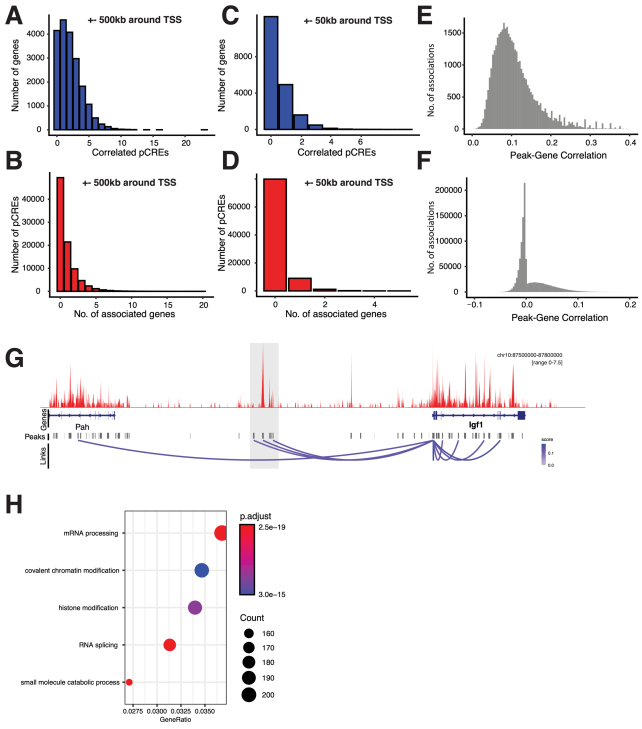
Supplementary Figure 2



Supplementary Figure 2: easySHAREseq robustly separates cell types

(A) UMAP visualisation of merged and integrated scRNA-seq data. Nuclei are coloured according to their cell type.
(B) UMAP visualisation of merged and integrated scATAC-seq data. Nuclei are coloured according to their cell type.
(C) Fraction of cell types recovered relative to total cells
(D) Distribution of UMIs per cell split by cell type. Some cell types (e.g. LSECs) consistently yield less UMIs.
(E) Distribution of unique fragments per cell split by cell types. Some cell types (e.g. LSECs) consistently yield less fragments.
(F) WNN-UMAPs with cells coloured according to the mean expression strength of a given marker gene. Red circles indicate the position of the cell population showing elevated expression for this marker gene.

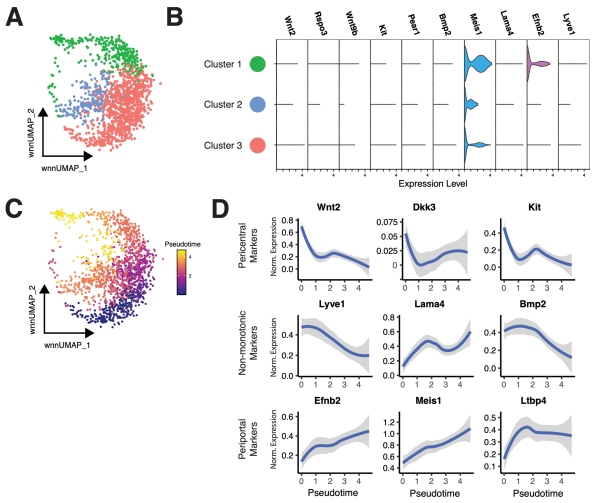
Supplementary Figure 3



Supplementary Figure 3: Summary of peak–gene correlations

(A) Number of significantly correlated pCREs ($P < 0.05$, FDR = 0.1) per gene, considering all peaks ±500kbp of the TSS
(B) Number of genes a given pCREs is significantly correlated with ($P < 0.05$, FDR = 0.1), considering all peaks ±500kbp of the TSS
(C) Number of significantly correlated pCREs ($P < 0.05$, FDR = 0.1) per gene, considering all peaks ±50kbp of the TSS
(D) Number of genes a given pCREs is significantly correlated with ($P < 0.05$, FDR = 0.1), considering all peaks ±50kbp of the TSS
(E) Histogram of Spearman correlations of all significant peak–gene correlations ($P < 0.05$)
(F) Histogram of Spearman correlations of all non-significant peak–gene correlations ($P > 0.05$)
(G) Aggregate scATAC-seq track of LSECs at the *Igf1* locus and its upstream region. Loops denote significantly correlated pCREs with *Igf1* and are coloured by their respective Spearman correlation. Shaded grey area denotes potentially LSEC-specific *cis*-regulatory element regulating *Igf1* expression.
(H) Gene Ontology enrichment analysis of genes whose associated pCREs are associated with five or more genes.

Supplementary Figure 4



Supplementary Figure 4: Investigation of LSEC zonation

(A) Subclustering LSECs reveals three distinct clusters.
(B) Comparison of marker gene expression across the three identified LSEC subclusters does not allow for fine-scale cell-type assignments.
(C) Subclustered LSECs coloured by pseudotime.
(D) Loess-Curve of marker gene expression of pericentral, non-monotonic and periportal marker genes along pseudotime.

Methods

Animal Model & Tissue preparation

Mice

All animal experimental procedures were carried out under the licence number EB 01-21M at Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, Germany. The procedures were reviewed and approved by the Regierungspräsidium Tübingen, Germany. Liver was collected from both male and female wild-type C57BL/6 and PWD/PhJ mice aged between 9 to 11 weeks.

Study design

From each strain, we generated easySHARE-seq libraries for one male and one female mice from each strain (four total). For each individual, we sequenced two sub-libraries, resulting in 8 easySHAREseq libraries.

Cell Culture

For the species-mixing experiment, HEK Cells were cultured in media containing DMEM/F-12 with GlutaMAX™ Supplement, 10% FBS and 1% Penicillin-Streptomycin (PenStrep) at 37°C and 5% CO_2. Cells were harvested on the day of the experiment by simply pipetting them off the plate and were then spun down for 5 min at 250G.
For the second cell line, murine OP9-DL4 cells were cultured in alpha-MEM medium containing 5% FBS and 1% PenStrep. On the day of the experiment, the cells were harvested by aspirating the media and adding 4 ml of Trypsin, followed by an incubation at 37°C for 5 min. Then, 5ml of media was added and cells were spun down for 5 min at 250G.
After counting both cell lines using TrypanBlue and the Evos Countess II, equal cell numbers were mixed.

Liver Nuclei

The liver was extracted, rinsed in HBSS, cut into small pieces, frozen in liquid nitrogen and stored in the freezer at -80 °C for a maximum of two weeks. On the day of the experiment, 1 ml of ice cold Lysis Solution (0.1% Triton-X 100, 1mM DTT, 10mMM Tris-HCl pH8, 0.1mM EDTA, 3mM $Mg(Ac)_2$, 3mM $CaCl_2$ and 0.32M sucrose) was added to the tube. The cell suspension was transferred to a pre-cooled Douncer and dounced 10x using Pestle A (loose) and 15x using Pestle B (tight). The solution was added to a thick wall ultracentrifuge tube on ice and topped up with 4ml ice cold Lysis Solution. Then 9 ml of Sucrose solution (10mM Tris-HCl pH8.0, 3mM $Mg(Ac)_2$, 3mM DTT, 1.8M sucrose) was carefully pipetted to the bottom of the tube to create a sucrose cushion. Samples were spun in a pre-cooled ultracentrifuge with a SW-28 rotor at 24,400rpm for 1.5 hours at 4 °C. Afterwards, all supernatant was carefully aspirated so as not to dislodge the pellet at the bottom and 1 ml ice cold DEPC-treated water supplemented with 10µl SUPERase & 15µl Recombinant RNase Inhibitor was added. Without resuspending, the tube was kept on ice for 20 min. The pellet was then resuspended by pipetting ~15 times slowly up and down followed by a 40 µm cell straining step. Counting of the nuclei using DAPI and the Evos Countess II was immediately followed up by fixation.

easySHARE-seq protocol

Preparing the barcoding oligonucleotides

There are two barcoding rounds in easySHARE-seq with 192 unique barcodes distributed across two 96-well plates in each round (see **Suppl. Table 1** for a full list of oligonucleotide sequences). Each barcode (BC) is pre-annealed as a DNA duplex for improved stability. The first round of barcodes contains two single-stranded linker sequences at its ends as well as a 5' phosphate group to ligate the different barcodes together. The first single-stranded overhang links the barcode to a complementary overhang at the 5' end of the cDNA molecule or transposed DNA molecule, which originates either from the RT primer or the Tn5 adapter. The second overhang (3bp) is used to ligate it to the second round of barcodes (**Fig.1B**). Each duplex needs to be annealed prior to cellular barcoding, preferably on the day of the experiment. No blocking oligos are needed.

The Round1 BC plates contain 10µl of 4µM duplexes in each well and Round2 BC plates contain 10µl of 6µM barcode duplexes in each well, all in Annealing Buffer (10mM Tris pH8.0, 1mM EDTA, 30mM KCl). Pre-aliquoted barcoding plates can be stored at -20 °C for at least three months. On the day of the experiment, the oligo plates were thawed and annealed by heating plates to 95 °C for 2 min, followed by cooling down the plates to 20 °C at a rate of -2 °C per minute. Finally, the plates were spun down. Until the annealed barcoding plates are needed, they should be kept on ice or in the fridge.

This barcoding scheme is very flexible and currently supports a throughput of ~350,000 cells (assuming 96 indexing primers) per experiment, limited only by sequencing cost and availability of indexing primer. The barcodes were designed to have at least a Hamming distance of 2. See Supplementary Notes for further details on the barcoding system and flexibility.

Tn5 *preparation*

Tn5 was expressed in-house as previously described [44]. Two differently loaded Tn5 are needed for easySHARE-seq, one for the tagmentation, loaded with an adapter for attaching the first barcodes (termed Tn5-B2S), and one for library preparation with a standard illumina sequencing adapter (termed Tn5-A-only). See Supplementary Table 1 for all sequences.

To assemble Tn5-B2S, two DNA duplexes were annealed: 20 µM Tn5-A oligo with 22 µM Tn5-reverse and 20 µM Tn5-B2S with 22 µM Tn5-reverse, all in 50 mM NaCl and 10mM Tris pH8.0. Oligos were annealed by heating the solution to 95 °C for 30 s and cooling it down to 20 °C at a rate of 2 °C/min. An equal volume of duplexes was pooled and then 200 µl of unassembled Tn5 was mixed with 16.5 µl of duplex mix. The Tn5 was then incubated at 37 °C for 1 hour, followed by 4 °C overnight. The Tn5 can then be stored at -20 °C. In our hands, Tn5 did not show a decrease in activity after 10 months of storage.

To assemble Tn5-A-only, 10 µM of Tn5-A and 10.5 µM Tn5-reverse was annealed using the same conditions as described above. Again, 200 µl of unassembled Tn5 was mixed with 16.5 µl of Tn5-A duplex and incubated at 37 °C for 1 hour, followed by 4 °C overnight. The Tn5 can then be stored for later and repeated use for more than 10 months at -20 °C.

We observed an increase in all Tn5 activity during the first months of storage, possibly due to continued transposome assembly in storage.

Fixation

One million liver nuclei ("cells" for short) were added to ice-cold PBS for 4 ml total. After mixing, 87 µl 16% formaldehyde solution (0.35%; for liver nuclei) or 25 µl 16% formaldehyde solution

(0.1%; for HEK and OP9 cells) was added and the suspension was mixed by pipetting up and down exactly 3 times with a P1000 pipette set to 700 µl. The suspension was incubated at room temperature for 10 min. Fixation was stopped by adding ice-cold Stop-Mix (224 µl 2.5M glycine, 200 µl 1M Tris-HCl pH8.0, 53 µl 7.5% BSA in PBS). The suspension was mixed exactly 3 times with a P1000 pipette set to 850 µl and incubated on ice for 3 min followed by a centrifugation at 500G for 5 min at 4°C. Supernatant was removed and the pellet was resuspended in 1 ml Nuclei Isolation Buffer (NIB; 10mM Tris pH8.0, 10mM NaCl, 2mM $MgCl_2$, 0.1% NP-40) and kept on ice for 3 min followed by straining the suspension with a 40 µm cell strainer. It was then spun down at 500G for 3 min at 4°C and re-suspended in ~100-200µl PBSi (1x PBS + 0.4 U/µl Recombinant RNaseInhibitor, 0.04% BSA, 0.2 U/µl SUPERase, freshly added), depending on the amount of input cells. Cells were then counted using DAPI and the Countess II and concentration was adjusted to 2M cells/ml using PBSi.

Tagmentation

In a typical easySHARE-seq experiment for this study, 8 tagmentation reactions with 10,000 cells each followed by 3 RT reactions were performed. This results in sequencing libraries for around 30,000 cells. To increase throughput, simply increase the amount of tagmentation and RT reactions accordingly. No adjustment is needed to the barcoding. Each tube and PCR strip until the step of Reverse Crosslinking was coated before use by rinsing it with PBS+0.5% BSA.

For each tagmentation reaction, 5 µl of 5X TAPS-Buffer, 0.25µl 10% Tween, 0.25µl 1% Digitonin, 3 µl PBS, 1 µl Recombinant RNaseInhibitor and 9µl of H2O was mixed. TAPS Buffer was made by first making a 1M TAPS stock solution in H_2O, followed by adjustment of the pH to 8.5 by titrating 10M NaOH. Then, 4.25ml H_2O, 500µl 1M TAPS pH8.5, 250µl 1M $MgCl_2$ and 5ml N-N-Di-Methyl-Formamide (DMF) was mixed on ice and in order. When adding DMF, the buffer heats up so it is important to be kept on ice. The resulting 5X TAPS-Buffer can then be stored at 4°C for short term use (1-2 months) or for long-term storage at -20°C (> 6 months).

Then, 5 µl of cell suspension at 2M cells/ml in PBSi was added to the tagmentation mix for each reaction, mixed thoroughly and finally 1.5µl of Tn5-B2S was added. The reaction was incubated on a shaker at 37°C for 30 min at 850 rpm. Afterwards, all reactions were pooled on ice into a pre-cooled 15ml tube. The reaction wells were washed with ~30 µl PBSi which was then added to the pooled suspension in order to maximize cell recovery. The suspension was then spun down at 500G for 3 min at 4°C. Supernatant was aspirated and the cells were washed with 200µl NIB followed by another centrifugation at 500G for 3 min at 4°C.

We only observed cell pellets when centrifuging after fixation and only when using cell lines as input material. Therefore, when aspirating supernatant at any step it is preferable to leave around 20-30µl liquid in the tube. Additionally, it is recommended to pipette gently at any step as to not damage and fracture the cells.

Reverse Transcription

As stated above, three tagmentation reactions were combined into one RT reaction. When increasing cells to more than 30,000 per RT reaction, we observed a steep drop in reaction efficiency.

The Master Mix for one RT reaction contained 3µl 100µM RT-primer, 2µl 10mM dNTPs, 6µl 5X MaximaH RT Buffer, 4.5µl 50% PEG6000, 1.5 µl H_2O, 1.5µl SUPERase and 1.66µl MaximaH RT. The RT primer contains a polyT tail, a 10bp UMI sequence, a biotin molecule and an adapter sequence used for ligating onto the first round of barcoding oligos.

The cell suspension was resuspended in 10µl NIB per RT reaction and added to the Master Mix for a total of 30µl. As PEG is present, it is necessary to pipette ~30 times up and down to ensure proper mixing. The RT reaction was performed in a PCR cycler with the following protocol: 52°C for 12 min; then 2 cycles of 8°C for 12s, 15°C for 45s, 20°C for 45s, 30°C for 30s, 42°C for 2min and 50°C for 3 min. Finally, the reaction was incubated at 52°C for 5 more minutes. All reactions were then pooled on ice into a pre-cooled and coated 15ml tube and the reaction wells were washed with ~40µl NIB, which was then added to the pooled cell suspension in order to maximise cell recovery. The suspension was then spun down at 500G for 3 min at 4°C. Supernatant was aspirated and the cells were washed in 150µl NIB and spun down again at 500G for 3min at 4°C. This washing step was repeated once more, followed by resuspension of the cells in 2ml Ligation Mix (400µl 10x T4-Buffer, 40µl 10% Tween-20, 1460µl Annealing Buffer and 100µl T4 DNA Ligase, added last).

Single-cell barcoding

Using a P20 pipette, 10µl of cell suspension in the ligation mix was added to each well of the two annealed Round1 BC plates, taking care as to not touch the liquid at the bottom of each well. The plates were then sealed, shaken gently by hand and quickly spun down (~ 8s) followed by an incubation on a shaker at 25°C for 30 min at 350 rpm. After 30 min, the cells from each well were pooled into a coated PCR strip using a P200 multichannel pipette set to 30µl. In order to pool, each row was pipetted up and down three times before adding the liquid to the PCR strip. After 8 columns were pooled into the strip, the suspension was transferred into a coated 5ml tube on ice. This process was repeated until both plates were pooled, taking care to aspirate most liquid from the plates. The cell suspension was then spun down for 3min at 500G at 4°C. Supernatant was aspirated and the cells were resuspended thoroughly in 2 ml new Ligation Mix. Now, 10µl of cell suspension was added into each well of the annealed Round2 barcoding plates using a P20 pipette, taking care as to not touch the liquid within each well. The plates were sealed, shaken gently by hand and spun down quickly followed by incubating them on a shaker at 25°C for 45 min at 350 rpm. The cells were then pooled again using the above described procedure into a new coated 15ml Tube. The cells were spun down at 500G for 3 min at 4°C. Supernatant was aspirated, the cells were washed with 150µl NIB and spun down again. Finally, the cells were resuspended in ~60µl NIB and counted. For counting, 5µl of cells were mixed with 5µl of NIB and 1x DAPI and counted on the Evos Countess II, taking the dilution into account. Sub-libraries of 3,500 cells were made and the volume was adjusted to 25µl by addition of NIB.

Using 3,500 cells results in a doublet rate of ~6.3%. The recovery rate of cells after sequencing depends on the input material (and QC thresholds), with cell lines recovering around 80% of input cells (~2,800-3,000 cells) and liver nuclei around 70% (~2,300-2,500 cells).

Reverse-Crosslinking

To each sub-library of 3,500 cells, 30µl 2x Reverse Crosslinking (RC) Buffer (0.4% SDS, 100mM NaCl, 100mM Tris pH8.0) as well as 5µl ProteinaseK was added. The sub-libraries were mixed and incubated on a shaker at 62°C for one hour at 800 rpm. Afterwards, they were transferred to a PCR cycler into a deep well module set to 62°C (lid to 80°C) for an additional hour. Afterwards, each sub-library was incubated at 80°C for 10 min and finally 5µl of 10% Tween-20 to quench the SDS and 35µl of NIB was added for a total volume of 100µl.

The lysates can be stored at this point at -20°C for at least two days, which greatly simplifies handling many sub-libraries at once. Longer storage has not been extensively tested.

Streptavidin Pull-Down

Each transcript contains a biotin molecule as the RT primers are biotinylated which is used to separate the scATAC-seq libraries from the scRNA-seq libraries. For each sublibrary, 50µl M280 Streptavidin beads were washed three times with 100µl B&W Buffer (5mM Tris pH8.0, 1M NaCl, 0.5mM EDTA) supplemented with 0.05% Tween-20, using a magnetic stand. Afterwards, the beads were resuspended in 100µl 2x B&W Buffer and added to the sublibrary, which were then shaken at 25°C for one hour at 900 rpm. Now all cDNA molecules are attached to the beads whereas transposed molecules are within the supernatant. The lysate was put on a magnetic stand to separate supernatant and beads.
It likely is possible to reduce the number of M280 beads in this step, significantly reducing overall costs. However, this has not been extensively tested.

scATAC-seq library preparation

The supernatant from each sub-library was cleaned up with a Qiagen MinElute Kit and eluted twice into 30µl 10mM Tris pH8.0 total. PCR Mix containing 10µl 5X Q5 Reaction Buffer, 1µl 10mM dNTPs, 2µl 10µM i7-TruSeq-long primer, 2µl 10µM Nextera N5XX Indexing primer, 4.5µl H2O and 0.5µl Q5 Polymerase was added (All Oligo sequences in **Suppl. Table 1**). Importantly, in order to distinguish the samples, each sub-library needs to be indexed with a different N5XX Indexing primer. The fragments were amplified with the following protocol: 72°C for 6 min, 98°C for 1 min, then cycles of 98°C for 10s, 66°C for 20s and 72°C for 45s followed by a final incubation at 72°C for 2 min. The number of PCR cycles strongly depends on input material (Liver: 17 PCR cycles, Cell Lines: 15 PCR cycles). The reactions were then cleaned up with custom size selection beads with 0.55X as upper cutoff and 1.4X as lower cutoff and eluted into 25µl 10mM Tris pH8.0. Libraries were quantified using the Qubit HS dsDNA Quantification Kit and run on the Agilent 2100 bioanalyzer with a High Sensitivity DNA Kit.

cDNA library preparation

The beads containing the cDNA molecules were washed three times with 200µl B&W Buffer supplemented with 0.05% Tween-20 before being resuspended in 100µl 10mM Tris ph8.0 and transferred into a new PCR strip. The strip was put on a magnet and the supernatant was aspirated. The beads were then resuspended in 50µl Template Switch Reaction Mix: 10µl 5X MaximaH RT Buffer, 2µl 100µM TS-oligo, 5µl 10mM dNTPs, 3µl Enzymatics RNaseIn, 15µl 50% PEG6000, 14µl H2O and 1.25µl MaximaH RT. The sample was mixed well and incubated at 25°C for 30 min followed by an incubation at 42°C for 90 min. The beads were then washed with 100µl 10mM Tris while the strip was on a magnet and resuspended in 60µl H_2O. To each well, 40µl PCR Mix was added containing 20µl 5X Q5 Reaction Buffer, 4µl 10µM i7-Tru-Seq-long primer, 4µl 10µM Nextera N5XX Indexing primer, 2µl 10mM dNTPs, 9µl H2O and 2µl Q5 Polymerase. The resulting mix can be split into two 50µl PCR reactions or run in one 100µl reaction. The PCR involved initial incubation at 98°C for 1 min followed by PCR cycles of 98°C for 10s, 66°C for 20s and 72°C for 3 min with a final incubation at 72°C for 5 min. Importantly, in order to distinguish the samples, each sub-library needs to be indexed with a different N5XX Indexing primer. The number of PCR cycles strongly depends on input material (Liver: 15 cycles, Cell lines: 13 cycles).
The PCR reactions were cleaned up with custom size selection beads using 0.7X as a lower cutoff (70µl) and eluted into 25µl 10mM Tris pH8.0. The cDNA libraries were quantified using the Qubit HS dsDNA Quantification Kit.

scRNA-seq library preparation

As the cDNA molecules are too long for sequencing (mean length > 700bp), they need to be shortened on one side. To achieve this, 25ng of each cDNA library was transferred to a new strip and volume was adjusted to 20µl using H_2O. Then 5µl 5X TAPS Buffer and 0.8µl Tn5-A-only was added and the sample was incubated at 55°C for 10 min. To stop the reaction, 25µl 1% SDS was added followed by another incubation at 55°C for 10 min. The sample was then cleaned up with custom size selection beads using a ratio of 1.3X and eluted into 30µl. Then 20µl PCR mix was added containing 10µl 5X Q5 reaction buffer, 1µl 10mM dNTPs, 2µl 10µM i7-Tru-Seq-long primer, 2µl 10µM Nextera N5XX Indexing primer (note: each sample needs to receive the **same** index primer as was used in the cDNA library preparation), 4.5µl H_2O and 0.5µl Q5 Polymerase. The PCR reaction was carried out with the following protocol: 72°C for 6 min, 98°C for 1 min, followed by 5 cycles of 98°C for 10s, 66°C for 20s and 72°C for 45s with a final incubation at 72°C for 2 min. Libraries were purified using custom size selection beads with a ratio of 0.5X as an upper cutoff and 0.8X as a lower cutoff. The final scRNA-seq libraries were quantified using the Qubit HS dsDNA Quantification Kit and run on the Agilent 2100 bioanalyzer with a High Sensitivity DNA Kit.

Sequencing

Both scATAC-seq and scRNA-seq libraries were sequenced simultaneously as they were indexed with different Index2 indices (N5XX). All libraries were sequenced on the Nova-seq 6000 platform (Illumina) using S4 2x150bp v1.5 kits (Read 1: 150 cycles, Index 1: 17 cycles, Index 2: 8 cycles, Read 2: 150 cycles). Libraries were partly multiplexed with standard Illumina sequencing libraries.

Custom Size selection beads

To make custom size selection beads, we washed 1ml of SpeedBeads on a magnetic stand in 1ml of 10mM Tris-HCl pH8.0 and re-suspended them in 50ml Bead Buffer (9g PEG8000, 7.3g NaCl, 500ul 1M Tris HCl pH8.0, 100ul 0.5M EDTA, add water to 50ml). The beads don't differ in their functionality from other commercially available ready-to-use size selection beads. They can be stored at 4°C for > 3 months.

Analysis

Gene annotations and Genomic variants

The reference genome and the Ensembl gene annotation of the C57BL/6J genome (mm10) were downloaded from Ensembl (Version GRCm38, release 102). Gene annotations for PWD/PhJ mice were downloaded from Ensembl. A consensus gene annotation set in mm10 coordinates was constructed by filtering for genes present in both gene annotations.

easySHARE-RNA-seq pre-processing

Fastq files were demultiplexed using a custom C-script, allowing one mismatch within each barcode segment. The reads were trimmed using cutadapt[48]. UMIs were then extracted from bases 1-10 in Read 2 using UMI-Tools[45] and added to the read name. Only reads with TTTTT at the bases 11-15 of Read 2 were kept (> 96%), allowing one mismatch. Lastly, the barcode was also moved to the read name.

Species-Mixing Experiments

RNA-seq reads were aligned to a composite hg38-mm10 genome using STAR[46]. The resulting bamfile was then filtered for uniquely mapping reads and reads mapping to chrM, chrY or unmapped scaffolds or containing unplaced barcodes were removed. Finally, the reads were deduplicated using UMItools[45]. ATAC-seq reads were also aligned to a composite genome using bwa[47]. Duplicates were removed with Picard tools and reads mapping to chrM, chrY or unmapped scaffolds were filtered out. Additionally, reads that were improperly paired or had an alignment quality < 30 were also removed.

The reads were then split depending on which genome they mapped to and reads per barcode were counted. Barcodes needed to be associated with at least 700 fragments and 500 UMIs in order to be considered a cell for the analysis. A barcode was considered a doublet when either the proportion of UMIs or fragments assigned to a species was less than 75%. This cutoff was chosen to mitigate possible mapping bias within the data.

easySHARE-RNA-seq processing and read alignment

We only used Read 1 for all our RNA-seq analyses as sequencing quality tends to drop after a polyT tail is sequenced in R2. Each sample was mapped to mm10 using the twopass mode in STAR[46] with the parameters --outFilterMultimapNmax 20 --outFilterMismatchNmax 15. We then processed the bamfiles further by moving the UMI and barcode from the read name to a bam flag, filtering out multimapping reads and reads without a definitive barcode. To determine if a read overlapped a transcript, we used featureCounts from the subread package[48]. UMI-Tools was used to collapse the UMIs of aligned reads, allowing for one mismatch and de-duplication of the reads. Finally, (single-cell) count matrices were created also using UMI-Tools.

easySHARE-ATAC-seq pre-processing and read alignment

Fastq files were demultiplexed using a custom C-script, allowing one mismatch within each barcode segment. The paired reads were trimmed using cutadapt[49] and the resulting reads were mapped to the mm10 genome using bwa mem[47]. Reads with alignment quality < Q30, unmapped, undetermined barcode, or mapped to mtDNA were discarded. Duplicates were removed using Picard tools. Open chromatin regions were called by subsampling the bamfiles from all samples to a common depth, merging them into a pooled bamfile and using the peak caller MACS2[50] with the parameters -nomodel -keep-dup -min-length 100. The count matrices as well as the FRiP score was generated using featureCounts from the Subread package[48] together with the tissue-specific peak set.

Filtering, Integration & Dimensional reduction of scRNAseq data

The count matrices were loaded into Seurat[51] and cells were then filtered for >200 detected genes, >500 UMIs and < 20.000 UMIs. The sub-libraries coming from the same experiment were then merged together and normalised. Merged experiments from the same species (one from male mouse, one from female mouse) were then integrated by first using SCTransform[52] to normalise the data, then finding common features between the two experiments using FindIntegrationAnchors() and finally integrated using IntegrateData(). Lastly, the integrated datasets from C57BL/6 and PWD/PhJ were again integrated using IntegrateData(). To visualise the data, we projected the cells into 2D space by UMAP using the first 30 principal components and identified clusters using FindClusters().

Filtering, Integration & Dimensional reduction of scATACseq data

Fragments per cell were counted using sinto and the resulting fragment file was loaded into Signac[53] alongside the count matrices and the peakset. We calculated basic QC statistics using base Signac and cells were then filtered for a FRiP score of at least 0.3, > 300 fragments, < 15.000 fragments, a TSS enrichment > 2 and a nucleosome signal < 4. Again, sublibraries coming from the same experiment were merged. We then integrated all four experiments (C57BL/6 & PWD/PhJ, one male & one female mouse each) by finding common features across datasets using FindIntegrationAnchors() using PCs 2:30 and then integrating the data using IntegrateEmbeddings(). To visualise the data, we projected the cells into 2D space by UMAP.

Weighted-Nearest-Neigbor (WNN) Analysis & Cell type identification

In order to use data from both modalities simultaneously, we created a multimodal Seurat object and used WNN[16] clustering to visualise and leverage both modalities for downstream analysis. Afterwards, we assigned cell cycle scores and excluded clusters consisting of nuclei solely in the G2M-phase (2 clusters, 121 nuclei total). Cell types were assigned via expression of previously known marker genes, which allows subsetting the data into cell types.

Calculating Peak–Gene Associations

Peak–gene associations were calculated following the framework described by Ma et al[13]. In short, Spearman correlation was calculated for every peak–gene pair within a +-500kb window around the TSS of the expressed gene. To obtain a background estimation, we used chromVAR[54] (*getBackgroundPeaks*()) to generate 100 background peaks matched in GC bias and chromatin accessibility but randomly distributed throughout the genome. We calculated the Spearman correlation between every background-gene comparison, resulting in a null distribution with known population mean and standard deviation. We then calculated the z-score for the peak–gene pair in question ((correlation - population mean)/ standard deviation) and used a one-sided z-test to determine the p-value. This functionality is also implemented in Signac under the function *LinkPeaks*(). Increasing the number of background peaks to 200, 350 or 500 for each peak–gene pair does not impact the results (*data not shown*).

Analysis of LSEC zonation markers

To analyse gene expression and chromatin accessibility along LSEC zonation, we subsetted our data for LSECs only, extracted expression values and wnnUMAP coordinates and binned the data along the wnnUMAP_2 axis into 10 equal sized bins. We then calculated the mean expression/accessibility for each gene/peak in each bin, excluding cells that contained a zero count. To identify novel marker genes, we excluded genes with low expression and calculated the moving average (for three bins) across the bins. We then required the moving average to continuously decrease (for pericentral marker genes) or increase (for periportal marker genes), allowing two exceptions. Lastly, we divided the means for each gene by their maximum to normalise the values. Identification of *cis*-regulatory elements displaying zonation effects had equal requirements.

Imputation of pseudotime was performed in Monocle3[55] with standard parameters. Gene expression was smoothed over both bins and pseudotime (separately) with local polynomial regression fitting (loess).

Gene Ontology Analysis

Gene Ontology Analysis was done using the R package clusterProfiler[56] with standard parameters.

Data Availability

All data can be accessed using the accession number GSE256434. All code used in data analysis is available at https://github.com/vosoltys/easySHARE_seq.git.

Acknowledgements

We thank members of the Chan and Jones lab for helpful discussions and critical reading of the manuscript. We are very grateful to Arnar Breevoort and Alex Pollen for sharing tissue preparation protocols and a very helpful research visit. We thank Sinja Mattes and all animal care takers at the Friedrich Miescher Laboratory for their work. We also thank the Genome Center in the Max Planck Institute for Biology Tübingen for providing support. The OP9-DL4 cells were a kind gift from Juan Carlos Zúñiga-Pflücker. M.P. is supported by an International Max Planck Research School fellowship. M.K. and Y.F.C. were supported European Research Council Starting Grant 639096 “HybridMiX" and Proof-of-Concept Grant 101069216 “Haplotagging”. The research was supported by the Max Planck Society.

Author Contributions

V.S. and Y.F.C. designed the experiments. V.S. and M.P. developed the barcoding framework for easySHAREseq. V.S. developed the rest of the protocol and performed experiments. V.S. performed the computational analyses advised by Y.F.C. V.S. drafted the manuscript. M.P., D.S., M.K. and Y.F.C. helped with experimental or computational support. All authors reviewed the manuscript. Y.F.C. directed the study with input from all authors.

Declaration of Interest

The authors declare no competing interests.

References

1. Anderson, E., Devenney, P. S., Hill, R. E. & Lettice, L. A. Mapping the Shh long-range regulatory domain. *Development* **141**, 3934–3943 (2014).
2. Nord, A. S. *et al.* Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell* **155**, 1521–1531 (2013).
3. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).
4. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).
5. Zhang, C., Macchi, F., Magnani, E. & Sadler, K. C. Chromatin states shaped by an epigenetic code confer regenerative potential to the mouse liver. *Nat Commun* **12**, 4110 (2021).
6. Lara-Astiaso, D. *et al.* Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
7. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
8. Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp Mol Med* **52**, 1419–1427 (2020).
9. Martin, B. K. *et al.* Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nat Protoc* **18**, 188–207 (2023).
10. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
11. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
12. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* **37**, 1452–1457 (2019).
13. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).
14. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 31 (2020).
15. Aizarani, N. *et al.* A Human Liver Cell Atlas reveals Heterogeneity and Epithelial Progenitors. *Nature* **572**, 199–204 (2019).
16. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
17. Su, Q. *et al.* Single-cell RNA transcriptome landscape of hepatocytes and non-parenchymal cells in healthy and NAFLD mouse liver. *iScience* **24**, 103233 (2021).
18. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e17 (2018).
19. Chen, X. *et al.* Structural insights into preinitiation complex assembly on core promoters. *Science* **372**, eaba8490 (2021).
20. Winkler, M. *et al.* Endothelial GATA4 controls liver fibrosis and regeneration by preventing a pathogenic switch in angiocrine signaling. *J Hepatol* **74**, 380–393 (2021).
21. Géraud, C. *et al.* GATA4-dependent organ-specific endothelial differentiation controls liver development and embryonic hematopoiesis. *J Clin Invest* **127**, 1099–1114.
22. Lara-Diaz, V. *et al.* IGF-1 modulates gene expression of proteins involved in inflammation, cytoskeleton, and liver architecture. *J Physiol Biochem* **73**, 245–258 (2017).

23. Baratta, J. L. *et al.* Cellular Organization of Normal Mouse Liver: A Histological, Quantitative Immunocytochemical, and Fine Structural Analysis. *Histochem Cell Biol* **131**, 713–726 (2009).
24. Jungermann, K. & Kietzmann, T. Zonation of parenchymal and nonparenchymal metabolism in liver. *Annu Rev Nutr* **16**, 179–203 (1996).
25. Braeuning, A. *et al.* Differential gene expression in periportal and perivenous mouse hepatocytes. *The FEBS Journal* **273**, 5051–5061 (2006).
26. Planas-Paz, L. *et al.* The RSPO–LGR4/5–ZNRF3/RNF43 module controls liver zonation and size. *Nat Cell Biol* **18**, 467–479 (2016).
27. Wang, B., Zhao, L., Fish, M., Logan, C. Y. & Nusse, R. Self-renewing diploid Axin2+ cells fuel homeostatic renewal of the liver. *Nature* **524**, 180–185 (2015).
28. Halpern, K. B. *et al.* Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat Biotechnol* **36**, 962–970 (2018).
29. Knolle, P. A. & Wohlleber, D. Immunological functions of liver sinusoidal endothelial cells. *Cell Mol Immunol* **13**, 347–353 (2016).
30. Smedsrød, B. Clearance function of scavenger endothelial cells. *Comparative Hepatology* **3**, S22 (2004).
31. Rafii, S., Butler, J. M. & Ding, B.-S. Angiocrine functions of organ-specific endothelial cells. *Nature* **529**, 316–325 (2016).
32. Theilmann, A. L. *et al.* Endothelial BMPR2 Loss Drives a Proliferative Response to BMP (Bone Morphogenetic Protein) 9 via Prolonged Canonical Signaling. *Arteriosclerosis, Thrombosis, and Vascular Biology* **40**, 2605–2618 (2020).
33. Russell, K. S., Stern, D. F., Polverini, P. J. & Bender, J. R. Neuregulin activation of ErbB receptors in vascular endothelium leads to angiogenesis. *American Journal of Physiology-Heart and Circulatory Physiology* **277**, H2205–H2211 (1999).
34. Vihanto, M. M. *et al.* Hypoxia up-regulates expression of Eph receptors and ephrins in mouse skin. *FASEB J* **19**, 1689–1691 (2005).
35. Shen, Z. *et al.* Delta-Like Ligand 4 Modulates Liver Damage by Down-Regulating Chemokine Expression. *Am J Pathol* **186**, 1874–1889 (2016).
36. Zellmer, S. *et al.* Transcription factors ETF, E2F, and SP-1 are involved in cytokine-independent proliferation of murine hepatocytes. *Hepatology* **52**, 2127–2136 (2010).
37. Dong, X. C. *et al.* Inactivation of Hepatic Foxo1 by Insulin Signaling Is Required for Adaptive Nutrient Homeostasis and Endocrine Growth Regulation. *Cell Metabolism* **8**, 65–76 (2008).
38. Wang, X., Yu, Y., Xie, H.-B., Shen, T. & Zhu, Q.-X. Complement regulatory protein CD59a plays a protective role in immune liver injury of trichloroethylene-sensitized BALB/c mice. *Ecotoxicology and Environmental Safety* **172**, 105–113 (2019).
39. Ren, H. *et al.* Sirtuin 2 Prevents Liver Steatosis and Metabolic Disorders by Deacetylation of Hepatocyte Nuclear Factor 4α. *Hepatology* **74**, 723 (2021).
40. Miyao, M. *et al.* Pivotal role of liver sinusoidal endothelial cells in NAFLD/NASH progression. *Laboratory Investigation* **95**, 1130–1144 (2015).
41. Su, T. *et al.* Single-Cell Transcriptomics Reveals Zone-Specific Alterations of Liver Sinusoidal Endothelial Cells in Cirrhosis. *Cellular and Molecular Gastroenterology and Hepatology* **11**, 1139–1161 (2021).
42. Nikopoulou, C. *et al.* Spatial and single-cell profiling of the metabolome, transcriptome and epigenome of the aging mouse liver. *Nat Aging* **3**, 1430–1445 (2023).
43. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).

44. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
45. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491–499 (2017).
46. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
49. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
50. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
51. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).
52. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019).
53. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333–1341 (2021).
54. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975–978 (2017).
55. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).
56. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).

Discussion

TCR repertoire analysis has come a long way from analyzing a small collection of TCRβ chains to evaluating large portions of an individuals paired αβ-TCR repertoire. In the latter case, the repertoire data presented in this thesis is likely to represent the largest paired TCR dataset analyzed so far. Through the development of CITR-seq we were no longer limited by financial or technical constraints but rather by the amount of available input material (e.g. about 20mg of spleen tissue collected from SPRET mice). We reported remarkable TCR repertoire diversity and showed that in the joined TCR repertoire of all 32 individual mice consisting of roughly 5 million paired TCRs, about 95% of clonotypes were unique. Considering the age and husbandry conditions of the studied mice, it is fair to assume that the overwhelming majority of the sampled $CD8^+$ T cells were antigen-inexperienced, naïve T cells. Our primary focus in the presented study was to investigate how TCR repertoires are shaped by genetic factors, beyond the classical view of diversity generation by stochastic effects. We leveraged the distinct genetic backgrounds of wild-derived inbred mouse species and their F1 hybrids to provide evidence that a) genetic factors have a significant impact on the generative biases of particular V-J pairs during V(D)J recombination, that b) thymic selection can introduce additional significant biases to gene segment usage in a MHC-dependent manner and that c) the genetic background impacts the total repertoire diversity and extent of clonotype sharing between individuals. To the best of our knowledge, the respective wild-caught inbred mouse strains have not been systematically analyzed with respect to their TCR repertoires. Crucially, these mice can be used to evaluate TCR repertoire characteristics in the context of an evolutionary divergence time that is much broader than in any human study. While allelic MHC diversity in this setup is extremely sparse compared to outbred population, the often-overlooked allelic diversity of TCR loci is presumably well captured within the different species. Collectively, our findings provided a comprehensive view on species-specific TCR repertoire generation dynamics. In the following section I will discuss how our findings can contribute to address some of the long-standing debates in the field of TCR biology.

On the co-evolution of TCRs and MHCs

A mandatory prerequisite for all functional TCRs is their ability to form complexes with antigen presenting MHCs [10]. Structural analysis of TCR-MHCs has shown that the topology of the complex is conserved in a way that positions specific germline-encoded V gene segment regions of the TCR in proximity of the MHC molecule (CDR1 and CDR2) and its bound antigen (CDR3) [25]. It is therefore tempting to conclude that this very specific interaction, representing a key difference to the antigen binding of BCRs and antibodies, is under evolutionary selective pressure to favor those (V) gene segments that are capable of binding MHCs. The diversity generated by germline-independent somatic rearrangements of the CDR3 region makes it unlikely that these sequences are driving the MHC binding of TCRs. Therefore, the focus has mostly been set on CDR1 and CDR2 sequences. While those sequences are relatively conserved in length compared to the corresponding sequences in immunoglobulins [223], no highly conserved amino acid motifs encoding for the MHC binding capability have been identified to date [89]. Therefore, the alternative hypothesis emerged that MHC restriction is not germline-encoded but rather positive selection simply selects those TCRs capable of binding the MHC complexes. This hypothesis is challenged by the conserved TCR-MHC topology and studies that showed high frequencies of MHC-binding TCRs in pre-selection repertoires [224]. In essence, the debate on whether or not TCR-MHC binding is a co-evolutionary process has not been settled to date.

In this context several interesting findings can be retrieved from the present study, especially from the analysis of V(D)J gene segment usage in F1 hybrids. We showed that thymic selection can (although rarely) introduce drastic allele specific biases to Vβ gene segment usage in the TCR repertoire of mature T cells. These biases seem to be independent of the CDR3 sequence since they are not correlated with the usage of particular Vβ-Jβ combinations but rather the respective Vβ gene segment allele is rejected categorically (e.g. SPRET *Trbv13-2* allele in BL6xSPRET F1 hybrids). Based on this observation we concluded that this categorical rejection is unlikely to be caused by TCR self-reactivity as evaluated during negative selection but is likely the consequence of inappropriate MHC binding ability as determined during positive selection. More specifically, we propose that in the above example, the MHC-affinity of SPRET *Trbv13-2*

V gene segment, including TCRs is too strong and therefore the respective T cells receive apoptotic signals and are depleted from the repertoire. This would also mean that allelic bias of V gene segments is not a consequence of the mostly stochastic recombination process but has genetically encoded origin. While mapping of those potential genetic variants and validation in functional assays is beyond the scope of this project, we note that we frequently find SNPs in the CDR1 and CDR2 sequences of Vβ genes showing allele-specific biases in post selection repertoires. In line with this hypothesis is the observation that allelic biases emerging during thymic selection are almost absent in J gene segments, which are more relevant in peptide recognition compared to MHC binding. One might argue that because of the distinct MHC-haplotypes in those F1 hybrids of inbred mice, the above example represents a unique MHC-haplotype dependent case. Strikingly across all F1 hybrids we also observed shared Vβ gene segment rejection patterns. For example, while all F1 hybrids frequently recombine *Trbv12* family members, all TCRs containing these V gene segments are effectively rejected during thymic selection. Following the arguments outlined above, this is another strong example for germline encoded TCR-MHC incompatibilities. Nonetheless, to fully test this hypothesis, one would need to explore categorical rejection of V gene segments in an even broader MHC-haplotype context, for instance in F1 hybrids of a collection of wild-caught *Mus musculus domesticus* and *Mus spretus* individuals.

A separate question that arises from our results is why such examples have not yet been observed in human TCR repertoire studies. One possible explanation is that the F1 hybrids of wild-caught inbred mouse species evaluated in this study share their latest common ancestor between 0.5 and 3 million years ago. This evolutionary time span is much larger compared to estimates of the latest common ancestor of all living humans [225]. In general, categorical TCR-MHC incompatibilities in F1 hybrid mice were rare, limited mostly to Vβ gene segments and increased in frequency with increasing evolutionary divergence of the parental species. It is therefore entirely possible that V(D)J gene segment alleles and MHC-haplotypes across various human populations have not diverged enough to frequently create such incompatibilities (or they are so rare that those that exist have not been described yet). *Marrack et al.* argued, that the reason for the absence of a conserved germline-encoded MHC-binding motif in TCRs could be the

demand for a certain level of flexibility in TCR-MHC binding to compensate for CDR3 motif (length) diversity [226]. Further, while V gene segments with the ability to bind MHCs might be evolutionarily favored, those that exhibit too strong MHC affinity are depleted from the repertoire during thymic selection. Collectively, this might lead to a situation in which the decisive amino acid residues encoding MHC-binding abilities are relatively masked in the TCR repertoires of mature T cells.

In summary, the data presented here is more in line with the hypothesis that there is some form of genetically encoded ability of TCRs to bind MHC complexes. Considering the presented examples, it is hard to imagine that positive selection performs opportunistic selection of a small set of TCRs that happen to bind MHCs with just the right affinity from a sea of TCRs that have no pre-encoded MHC affinity at all. We see that categorical rejection of gene segments is limited to Vβ and happens in a strictly MHC-dependent manner.

The effect of MHC heterozygosity on the TCR repertoire

In the longstanding debate on whether heterozygous MHC loci confer a fitness advantage or disadvantage, TCR repertoire diversity can be used to provide evidence for one or the other. On the one hand, the TCR depletion hypothesis states that depletion of autoreactive TCRs is increasing with increasing levels of MHC heterozygosity leading to an effective size-reduction of the TCR repertoire ([227] and reviewed here [228]). Recently, *Migalska et al.* reported that the hypothesis can be applied to MHC class I but not class II, potentially caused by the ability of autoreactive $CD4^+$ T cells to adapt to a regulatory T cell fate rather than being depleted during negative selection [229]. In this context, it is also remarkable that across several species the total count of different intra-individual MHC molecules is relatively small in the face of the immense population wide allelic diversity. Extreme examples have been reported for the polyploid clawed toad (Xenopus) in which all but a single MHC locus are silenced, however these observations have not been associated with TCR diversity [230].

On the other hand, several lines of evidence support the heterozygote advantage hypothesis, stating that a diverse set of MHC alleles leads to the presentation of a broader immunopeptidome [231, 232]. Heterozygous MHC allele states in these studies are often

evaluated by long term reproductive success, motivated by several observations that intermediate levels of MHC heterozygosity are most frequent in outbred populations [233] and seem to be the preference in mate choice experiments [234, 235]. A clear example of TCR-related MHC heterozygosity advantage has been reported in two coisogenic mouse strains with significantly different survival rates following pathogen exposure [236]. Empirical evidence for increased TCR diversity in heterozygous HLA type I individuals has also been shown in a study of 666 humans of diverse origins [21]. It is now widely accepted that the fundamental mechanisms supporting one or the other hypothesis jointly affect the TCR repertoire. Mathematical models, have provided empirical evidence that the trade-off between enhanced antigen-presentation and increased rates of self-reactive T cell depletion in individuals with varying extent of MHC-heterozygosity, favors more intra-individual MHC diversity than observed in humans [237]. If TCR diversity reduction is not a limiting factor for intra-individual MHC diversity, then why is it still frequently observed in various species and populations? One potential explanation is the reported high level of TCR cross-reactivity to different antigens [238]. Cross-reactivity is not limited to pathogen-derived antigens but can also invoke autoimmune responses [239]. Therefore, increasing the number of different intra-individual MHC molecules can potentially also increase the risk of triggering autoimmunity through TCR cross-reactivity. The inbred mice analyzed in the context of this study are nowhere close to resemble the MHC diversity of an outbred population, which poses a clear limitation to answer the above questions. However, due to the traces of co-evolution between TCRs and MHCs discussed above, it is likely that the sparse selection of MHC alleles can still provide insides in heterozygous MHC combinations and their effect on repertoire diversity. In general, all F1 hybrids showed greater (paired) TCR diversity (number of unique CDR3 sequences) than both respective parents, which is in line with the heterozygote advantage hypothesis. However, only for the TCRα chain the diversity increase was correlated with increasing evolutionary divergence of the respective parental individuals. In contrast, for TCRβ chains as well as paired αβ-TCR chains, we saw the smallest repertoire diversity increase in BL6xSPRET mice in which the respective parents shared to most ancestral common ancestor. Considering that this particular cross of parental lines is close to the reported speciation barrier of mice [240], it is likely that crosses resulting in F1 hybrids

with decreased total TCR diversity relative to their parents fall beyond the species barrier. Importantly, we provide evidence (discussed in the previous section) that this diversity reduction is unlikely to be caused by depletion of self-reactive TCRs (as proposed by the TCR depletion hypothesis) but rather explained by the increased likelihood of insufficient TCR-MHC binding characteristics independent of the recognized antigen. We therefore conclude that, apart from potentially increased depletion of autoreactive TCRs, MHC haplotypes consisting of two highly divergent alleles have an increased chance to also limit the TCR repertoire diversity through categorical rejection of particular V gene segments. Regardless of this, even the joint TCR diversity reduction caused by both effects does apparently not outcompete TCR repertoire diversity increase caused by the presentation of a larger immunopeptidome in MHC heterozygous individuals.

In any case, TCRs need to function in combination with an extremely broad set of potential MHC molecules arising from haplotype diversity across a population. This diversity has emerged from variance in local pathogen exposures and essentially represents a case of host-pathogen co-evolution [241]. To some extent the required TCR repertoire flexibility might be established by the immense excess of unique TCRs in the theoretical repertoire relative to the realized repertoire. This ensures that even with approximately 95% of TCRs that fail to pass thymic selection, the mature repertoire is still sufficient to mount effective immune responses against most pathogens. Different studies have shown that a reduction in TCR diversity can be associated with impaired immune responses [242, 243]. However, these cases evaluate diversity reductions that are much more severe than the reduction caused by MHC heterozygosity reported in our as well as other studies.

Sharing of TCRs – How to become public

At a first glance, the sharing of identical TCRs across several individuals or even within entire populations might seem extremely unlikely, given the gigantic diversity in TCR repertoires. Yet, shared motifs are frequently observed across various TCR datasets generated from different species (summarized by [244]). Likewise, in the CITR-seq data presented here, we identified ~260.000 (36.7% of all unique motifs) CDR3α and ~470.000 (27.2% of all unique motifs) CDR3β single-chain amino acid motifs that were shared by at least two individuals. One mechanism proposed to explain the frequent observation of

shared CDR3 motifs is convergent recombination, stating that multiple V(D)J recombination events can converge to result in identical CDR3 amino acid sequences [122]. We provided evidence that TCR gene segment loci have been expanded by means of gene duplications, resulting in multiplied gene segments with high sequence identity (e.g., in the Vα cluster). As indicated by the difference in the number of gene family members, these gene duplications did not affect all V(D)J gene segments to a similar extent. Consequently, the high sequence identity between gene segments assigned to families of varying size should result in overrepresentation of particular germline-contributed sequences in CDR3 motifs. Interestingly, we observed the highest Jaccard index of single-chain CDR3α sharing between individual CAST mice. Due to the absence of the recent major gene duplication of two-thirds of the Vα locus in CAST, those mice have about 70 fewer functional Vα gene segments and were also shown to lack entire Vα families (e.g., Trav16) in their repertoire. We therefore conclude that not only convergent recombination, but also significantly contracted V(D)J loci can lead to an increased rate of shared CDR3 motifs.

V and J gene segment germline sequences contribute different numbers of nucleotides to each CDR3 motif. Critically, the average number of contributed nucleotides is higher for J than for V gene segments. For instance, in CDR3β sequence of all BL6 CITR-seq samples, the average germline sequence contribution to the CDR3 motif was 14.03 nt from Vβ and 18.46 nt from Jβ. Due to the lack of D gene segments, this difference is even higher in TCRα chains. Consequently, J gene segment sequences extend further into the central region of CDR3 sequences, which arguably makes them more relevant for antigen recognition than V gene segment derived sequences. We noticed that many of the differentially abundant amino acid 4mers across the different species could be traced back to specific positional SNPs of J gene segments. For example, the 4mer “NAET” was significantly more abundant in central BL6 CDR3β amino acid motifs relative to CAST (log2FC 6.18, Wald-test *adj. p-value* < 0.001). This 4mer was almost exclusively found to be derived from *Trbj2-3* in BL6 that contains a non-synonymous SNP relative to the CAST *Trbj2-3* (E3A). Due to the specific location at the 5’ end of the J gene segment germline sequence (3rd amino acid), this motif was frequently unaffected by nucleotide deletions during gene segment junction fusion, yet it was located in the central CDR3 sequence

and therefore likely contributes to antigen specificity of the TCR. Because the number of Jβ genes is relatively small (12 functional genes in all species), such SNP-related motif differences can affect large portions of CDR3β motif repertoire and consequently increase the likelihood of intra-species CDR3 motif sharing. Allelic sequence variation in V(D)J gene segments in humans has also been shown to impact immune responses [245]. However, this study focused on V gene segment polymorphisms outside the 3' coding end (contributing to the CDR3 motif), as a potential reason for the emergence of a disease associated public motif. Collectively, polymorphisms in TCR V(D)J loci have gained very little attention so far. Nonetheless, depending on their precise location (e.g. in the segment coding ends represented in CDR3 motifs), they potentially contribute to the increased frequency of intra-species public motifs.

Apart from the contribution of germline sequences to CDR3 motifs, nucleotide additions by terminal deoxynucleotidyl transferase (TdT) significantly increase CDR3 motif diversity [246] and polymorphisms in its coding sequence have been shown to alter the number of inserted nucleotides [247]. Accordingly, CDR3 motifs of particular lengths should also be present at higher frequencies given their higher likelihood of generation resulting from the dynamics of nucleotide deletions and insertions [248]. In the presented CITR-seq data, unique amino acid motif length is normally distributed with a cross-species average of 14.10 amino acids in CDR3α and 14.35 amino acids in CDR3β motifs. In line with the previously suggested motif length reduction, we observe a slight decrease of mean amino acid motif length in public CDR3 sequences (-1.4% CDR3α and -2.2% CDR3β). Strikingly, the 1,696 CDR3α and 644 CDR3β amino acid sequences that were present in every single of the 32 analyzed TCR repertoires showed a mean decrease in motif length of more than one amino acid relative to the total average. Nucleotide insertions are also biased with respect to the identity of added nucleotides, depending on the respective junction site and gene segment coding ends they are added to [249]. Taking into account the marked differences we observed in gene segment usage of post-selection repertoires across mouse species, those are likely to result in different abundances of coding-end sequences which in turn will impact nucleotide insertions by TdT and ultimately the likelihood of generating public CDR3 sequences. A common question that arises in this context is the exact stage of TCR generation at which publicness is established. The

impact of thymic selection on shaping the frequency of public CDR3 motifs has been addressed by several studies often generating contradicting results. Some studies provide evidence that publicness is established by biases of recombination frequencies of V(D)J gene segments, which is not altered in the subsequent thymic selection [244, 250]. On the other hand, there is evidence that thymic-selection represents a diversity bottle-neck and the decreased diversity of post-selection repertoires potentially increases the chance of CDR3 motif sharing [251]. In our CITR-seq data we observed significant differences in V(D)J segment usage in pre-selection repertoires. These usage differences, combined with the previously discussed difference in gene family sizes, will inherently lead to the overrepresentation of specific sequences before thymic selection. Utilizing F1 hybrid mice we provided evidence that most of the usage biases arise through *cis*-acting factors since parental usage frequencies were frequently recapitulated by allelic bias of gene segment usage in F1 hybrids. Accordingly, *trans*-acting factors, such as chromatin remodelers that make gene segments accessible for the recombination machinery, seem to be less relevant in the establishment of biased gene segment usage. Although not evaluated in this study, we propose that polymorphisms in RSS sequences targeted by the *Rag*-complex could alter their likelihood of being part of an recombination event, turning them into potent *cis*-regulatory elements (also reported here [252]). We see that the majority of the established gene segment usage biases persist in post-selection repertoires. However, we report strong exceptions especially for Vβ genes. By leveraging the distinct MHC-haplotype background of inbred mice and their F1 hybrids we could show that particular Vβ genes are almost completely rejected during thymic (positive) selection despite their frequent incorporation during V(D)J recombination. This has important implications for the generation of public CDR3 motifs. The set of shared CDR3 sequences is depleted of those motifs that were rejected during thymic selection. For example, in all F1 species hybrids, Vβ genes of the *Trbv12* family are significantly reduced in post- versus pre-selection repertoires. Consequently, shared CDR3β motifs are depleted from sequences originated from *Trbv12* coding-ends. In contrast, shared motifs across parental lines (that did not show thymic rejection of *Trbv12* genes) did not show depletion of those sequences. While this only affected a small set of (mostly V) gene segments, we therefore propose, that publicness within a set of repertoires should always

be evaluated in the context of the represented MHC-haplotypes. Another important finding of our study was that sharing of identical paired CDR3αβ sequences is significantly higher in inbred individuals that share the same genotype than in unrelated individuals. This underscores the importance of genetic factors in generating public CDR3 motifs. We propose that the collection of these genetic factors includes differences in functional V(D)J gene segment numbers, biases in gene segment usage, as well as MHC-haplotype dependent characteristics of thymic selection. In conclusion, our results are in line with previous studies reporting that publicness is established by a combination of convergent recombination and *cis*-factor mediated biases in gene segment usage frequencies. However, we note that, depending on the present MHC-haplotypes, the degree of TCR sharing can vary. While our results indicated that rejection of particular V gene segments is likely to reduce the degree of motif sharing, it is also generally possible that diversity reduction by negative thymic selections increases the likelihood of CDR3 motif sharing.

Another critical aspect of the analysis of public TCRs is that the degree of TCR sharing between two individuals is inherently biased by the union size of the sampled repertoires. While indices such as the Jaccard index can be used to compensate for this bias in multiple comparisons across several repertoires, absolute numbers of shared CDR3 amino acid sequences need to be evaluated with caution. Further, clone-size distributions are important indicators for the cause of TCR sharing. The presented CITR-seq data consists of total $CD8^+$ T cells and therefore includes naïve as well as memory T cell populations, which could not be distinguished. High-frequency clones that are shared across individuals might originate from memory T cell subsets resulting from previous clonal expansion of T cells with identical TCRs in response to the encounter of common antigens. It is therefore likely that exposure to common pathogens across individuals leads to the accumulation of public TCRs in their T cell memory compartments. It is important to distinguish those from public TCRs that can be found among antigen-inexperienced T cells as those provide evidence for antigen-exposure independent generation of public TCRs. *Mark et al.* suggested that the extend of CDR3 sequencing sharing is higher than previously expected, but many of the shared sequences are “hidden” at low frequencies and are only recognized following antigen-exposure [253]. Considering the immense throughput and ability to pair αβ-TCRs, CITR-seq represents

an exceptional tool to identify those “hidden” shared CDR3 sequences when specifically applied to naïve T cells.

To date, most studies of public TCRs have identified shared CDR3 amino acid motifs in the context of responses to common pathogens such as cytomegalovirus (CMV) and Epstein-Barr virus (EBV) [254-256]. Taking the latter as an example, it was shown that EBV viruses have evolved remarkable host specificity and have been infecting humans and their ancestors for approximately 80 million years [257]. Today, large portions of the human population are persistently infected by those pathogens and thus, loss and gain of particular MHC alleles should have favored those, that efficiently present peptides derived from these pathogens. Presumably the same is true for different sets of V(D)J gene segments showing varying affinity to MHC originated from the alleles under selective pressure. This poses the question whether public TCR motifs represent the outcome of the evolutionary arms race of host and pathogens that has shaped the collection of MHC alleles alongside V(D)J segments present in a population. In support of this hypothesis is the observation that many public motifs are a) mostly consisting of germline-contributed sequences [254] and b) are enriched for sequences with few or no random nucleotide insertions at the junction sites (discussed earlier). Public CDR3 motifs are not only abundant in the context of common pathogen infection but are also frequently found in the context of autoimmunity [258]. This phenomenon, is not exclusive to TCRs as shown by the identification of public autoreactive antibodies [259]. Mechanistically this again points towards the evolution of public TCRs as a consequence of frequent antigen encounters that are not limited to pathogen-derived antigens. We showed that decreased levels of CDR3 motif sharing does correlate with decreasing levels of genotype sharing and increasing evolutionary divergence. This is especially relevant as we see those effects in paired αβ-TCR repertoires of unprecedented scale and across multiple different inbred mouse species.

CITR-seq can be used in various research areas – an outlook

By developing CITR-seq we have been able to analyze the murine TCR repertoire at a scale that presumably captures a significant fraction of the available repertoire at the time of sampling. While we reported clear patterns of genotype-dependent V(D)J gene

segment usage and CDR3 motif sharing, the proportion of unique paired TCRs indicates that we mostly sampled a momentary snapshot of an individual's repertoire. Our T cell sampling method did not allow us to associate the evaluated TCRs with the diverse subclasses of $CD8^+$ T cells, however the observed clone size distributions indicate that most of the analyzed T cells were naïve T cells. As such, those T cells are valuable to analyze pathogen exposure independent repertoire generation and maintenance dynamics but are less suitable to investigate common immune responses to different pathogens or malignancies. Nonetheless, owing to its great flexibility, CITR-seq can be applied to much more specific research questions. For instance, the TCR repertoire is often evaluated in the context of human malignancies in pre- and post-treatment samples (reviewed here [260]). A frequently evaluated metric in these studies is the diversity of TCR repertoires that can potentially function as prognostic biomarker for treatment outcomes [261, 262]. Similarly, the ability to reconstruct a diverse TCR repertoire after hematopoietic stem cell transplantation has been associated with a decreased risk of cancer relapse [263]. It has been shown that tumor infiltrating lymphocytes frequently recognize a broad range of antigens that are not necessarily specific to the respective tumor [264]. Consequently, TCR sequencing methods like CITR-seq that provide a high dynamic range to detect TCRs of different clone sizes, are crucial to identify those tumor-neoantigen specific TCRs that might exist across different patients for specific types of cancer. Due to the high mutational load in various tumors at different progression stages, identification of common CDR3 motifs specific to tumor-antigens requires extremely large datasets that are often derived from heterogeneous sample collection databases [265]. The available datasets are often limited to single-chain TCR sequences (highly biased towards TCRβ chains) even though many studies have provided evidence that paired TCR information significantly enhances the ability to predict clinically relevant TCR epitopes [266-268]. When applied to different cancer-patient samples, CITR-seq can potentially be used to generate paired TCR repertoire data of sufficient size to confidently identify tumor-neoantigens present in various samples with significantly lower financial constraints compared to other methods. Additionally, supervised epitope-prediction algorithms (e.g. [181]) are dependent on large-scale TCR datasets with high αβ-TCR pairing accuracy, such as the data generated using CITR-seq.

Single-cell sequencing technologies are now rapidly moving towards simultaneously capturing multiple modalities (e.g. transcriptome and chromatin accessibility) in each individual cell. For instance, having access to the transcriptional profile of a T cell enables association with a specific T cell subpopulation and can therefore provide insights on whether an identified clinically relevant TCR is found among central memory, effector, or naïve T cells. Since, CITR-seq and easySHARE-seq are both based on identical single-cell barcoding strategies, it is generally possible to merge both technologies. Importantly, whole transcriptome analysis requires significantly higher per cell sequencing coverage (at least 200-fold higher compared to CITR-seq alone) and thus the extreme throughput provided by both methods would still be severely limited by financial constraints. With improvements in throughput and cost-effectiveness of current sequencing platforms, these constraints are likely to be less limiting in the future.

Closing Remarks

TCRs are arguably the most effective tool common to all vertebrates to survive in an environment of constantly varying pathogenic threats. The tremendous diversity of TCRs in the repertoire of an individual has fascinated researchers for decades but also severely complicates their analysis. Nowadays, high-throughput sequencing paired with specialized sequencing library preparation protocols, such as CITR-seq, allow us to evaluate and compare TCR repertoires at unprecedented scale. Access to such a wealth of TCR repertoire data is currently only used at a fraction of its full potential, partially because the ability to establish links between TCRs and their cognate antigens is still limited. With progress in computational prediction tools and the increased availability of experimentally validated TCR-antigen pairs, these limitations are likely to be overcome soon. At that point, one could imagine that TCRs turn into modular tools of future medicine that can be administered to patients suffering from various diseases. The current success of personalized cancer immunotherapies can offer a glimpse of T cell related therapies that might become the default in our future.

Acknowledgments

The cumulative work presented in this thesis represents years of exceptionally fruitful collaboration of the many colleagues and friends I met during my time as a PhD student at the Max Planck Institute for Biology in Tübingen. All of you have supported me throughout the bumpy road of my PhD and I would like to express my deep gratitude.

First and foremost, I would like to thank my supervisor, Frank. I consider myself extremely lucky to be given the chance to conduct the research that was precisely in line with my personal interest. You made this possible by providing the supervision, scientific training and resources and most importantly, the freedom to develop into the scientist I am today. I also want to thank the members of my TAC committee, Prof. Marja Timmermans, Prof. Cécile Gouttefangeas and Prof. Ralf Sommer, for providing the guidance and feedback that kept me on track to continuously progress with my projects. Additionally, I want to express my gratitude to Felicity Jones, who invoked many of the crucial ideas that resulted in the main findings of my project during our weekly lab meetings.

It has been a great privilege to peruse the journey of my PhD alongside my two amazing colleagues, Dingwen and Volker. Both of you have helped so much to overcome both scientific and mental roadblocks and I could not have asked for better lab-mates to have by my side from day one all the way to the end of my PhD. I also want to thank Marek, my bench-neighbor for several years, from whom I could learn a sea of molecular biology technologies and tricks and who kept up my energy levels with a constant supply of sweets.

The entire floor of the Friedrich-Miescher Laboratory has been a remarkably supportive and friendly scientific home. I am thankful to all the great people that shared this scientific home with me over the past year:
Insa Hirschberg for introducing me to the FML tissue culture and keeping me company for the many hours we spent in this windowless chamber. Julia, for sharing my fear to press the start button of the 10x controller. I want to also thank all the past members of

the Chan and Jones group: Elena Avdievich, Cholpon Zhakshylikova, Melanie Kirch, João Castro, Stanley Neufeld, Enni Harjunmaa, Layla Hiramatsu and Sebastian Kick.

The scientific community of the Max Planck Institute extends far beyond the floor of the FML. I want to specifically thank several people of this great community: Aurora Panzera for helping me to tame Melody, Sinja Mattes for managing our mouse colony, the RST members for always having an open ear in difficult times and Rebecca Schwab for always standing up for Dingwen, Volker and me.

I want to thank all my friends and family for continuously listening to my complaints and for frequently pulling me out of my scientific tunnel.

Lastly, I want to thank Josephin. You have been my most significant constant throughout the entire PhD. Your endless support, encouragement and most importantly, love cannot be acknowledged enough.

Glossary

aa	Amino acid
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
BLOSUM	Block substitution matrix
bp	Base pair
°C	Celsius
CD	Cluster of differentiation
cDNA	Complementary DNA
CDR	Complementarity determining region
CITR-seq	Combinatorial indexing-based T cell receptor sequencing
CMV	Cytomegalovirus
CNV	Copy number variation
cTEC	Cortical thymic epithelial cells
DN	Double negative
DNA	Deoxyribonucleic acid
DP	Double positive
EBV	Epstein-Barr virus
e.g.	Exempli gratia or “for example”
FACS	Fluorescence activated cell sorting
FC	Fold change
gDNA	Genomic DNA
HLA	Human leukocyte antigen
IF	In-frame
LPS	Lipopolysaccharide
MACS	Magnetic-activated cell sorting
Mbp	Megabase pair
MHC	Major histocompatibility complex
min	Minutes
MMLV	Moloney murine leukemia virus
mRNA	Messenger RNA
mTEC	Medullary thymic epithelial cells
Mya	Million years ago
NHEJ	Non-homologous end joining
n.s.	Not significant
nSDI	Normalized Shannon diversity index
oligo	Oligonucleotide
OOF	Out-of-frame
PC	Principal component
PCR	Polymerase chain reaction
polyA	Poly-adenylated
PRR	Pattern recognition receptors
PTC	Premature termination codon
RNA	Ribonucleic acid
RNA-seq	RNA sequencing

RSS	Recombination signal sequences
RT	Reverse transcription
SNP	Single nucleotide polymorphism
SP	Single positive
SPLiT-seq	Split Pool Ligation-based Transcriptome sequencing
TCR	T cell receptor
TIR	Terminal inverted repeat
TLR	Toll-like receptor
TREC	T cell receptor excision cycle
UMI	Unique molecular identifier
V(D)J	*Variable*, *diversity* and *joining* gene segments

References

1. Anderson, K.V., G. Jurgens, and C. Nusslein-Volhard, *Establishment of dorsal-ventral polarity in the Drosophila embryo: genetic studies on the role of the Toll gene product.* Cell, 1985. **42**(3): p. 779-89.
2. Bowie, A.G. and L. Unterholzner, *Viral evasion and subversion of pattern-recognition receptor signalling.* Nat Rev Immunol, 2008. **8**(12): p. 911-22.
3. Guo, P., et al., *Dual nature of the adaptive immune system in lampreys.* Nature, 2009. **459**(7248): p. 796-801.
4. Pancer, Z., et al., *Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey.* Nature, 2004. **430**(6996): p. 174-80.
5. Cooper, M.D., R.D. Peterson, and R.A. Good, *Delineation of the Thymic and Bursal Lymphoid Systems in the Chicken.* Nature, 1965. **205**: p. 143-6.
6. Allison, J.P., B.W. Mcintyre, and D. Bloch, *Tumor-Specific Antigen of Murine T-Lymphoma Defined with Monoclonal-Antibody.* Journal of Immunology, 1982. **129**(5): p. 2293-2300.
7. Haskins, K., et al., *The Major Histocompatibility Complex-Restricted Antigen Receptor on T-Cells .1. Isolation with a Monoclonal-Antibody.* Journal of Experimental Medicine, 1983. **157**(4): p. 1149-1169.
8. Cantor, H. and E.A. Boyse, *Functional subclasses of T lymphocytes bearing different Ly antigens. II. Cooperation between subclasses of Ly+ cells in the generation of killer activity.* J Exp Med, 1975. **141**(6): p. 1390-9.
9. Kisielow, P., et al., *Ly Antigens as Markers for Functionally Distinct Subpopulations of Thymus-Derived Lymphocytes of Mouse.* Nature, 1975. **253**(5488): p. 219-220.
10. Zinkerna.Rm and P.C. Doherty, *Restriction of in-Vitro T Cell-Mediated Cytotoxicity in Lymphocytic Choriomeningitis within a Syngeneic or Semiallogeneic System.* Nature, 1974. **248**(5450): p. 701-702.
11. Matsumura, M., et al., *Emerging principles for the recognition of peptide antigens by MHC class I molecules.* Science, 1992. **257**(5072): p. 927-34.
12. Bouvier, M. and D.C. Wiley, *Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules.* Science, 1994. **265**(5170): p. 398-402.
13. Bevan, M.J., *Cross-priming for a secondary cytotoxic response to minor H antigens with H-2 congenic cells which do not cross-react in the cytotoxic assay.* J Exp Med, 1976. **143**(5): p. 1283-8.
14. Smyth, M.J. and J.A. Trapani, *Granzymes - Exogenous Proteinases That Induce Target-Cell Apoptosis.* Immunology Today, 1995. **16**(4): p. 202-206.
15. Chicz, R.M., et al., *Predominant Naturally Processed Peptides Bound to Hla-Dr1 Are Derived from Mhc-Related Molecules and Are Heterogeneous in Size.* Nature, 1992. **358**(6389): p. 764-768.
16. Luckheeram, R.V., et al., *CD4(+)T cells: differentiation and functions.* Clin Dev Immunol, 2012. **2012**: p. 925135.
17. Horton, R., et al., *Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project.* Immunogenetics, 2008. **60**(1): p. 1-18.
18. Beck, S., et al., *Complete sequence and gene map of a human major histocompatibility complex.* Nature, 1999. **401**(6756): p. 921-923.
19. Hughes, A.L. and M. Yeager, *Natural selection at major histocompatibility complex loci of vertebrates.* Annual Review of Genetics, 1998. **32**: p. 415-435.
20. Sharon, E., et al., *Genetic variation in MHC proteins is associated with T cell receptor expression biases.* Nature Genetics, 2016. **48**(9): p. 995-+.
21. Krishna, C., et al., *Genetic and environmental determinants of human TCR repertoire diversity.* Immun Ageing, 2020. **17**: p. 26.
22. Sim, B.C., et al., *Control of MHC restriction by TCR Valpha CDR1 and CDR2.* Science, 1996. **273**(5277): p. 963-6.
23. Chlewicki, L.K., et al., *High-affinity, peptide-specific T cell receptors can be generated by mutations in CDR1, CDR2 or CDR3.* Journal of Molecular Biology, 2005. **346**(1): p. 223-239.

24. Garboczi, D.N., et al., *Structure of the complex between human T-cell receptor, viral peptide and HLA-A2.* Nature, 1996. **384**(6605): p. 134-141.
25. Garcia, K.C., et al., *An alpha beta T cell receptor structure at 2.5 angstrom and its orientation in the TCR-MHC complex.* Science, 1996. **274**(5285): p. 209-219.
26. Reinherz, E.L., et al., *The crystal structure of a T cell receptor in complex with peptide and MHC class II.* Science, 1999. **286**(5446): p. 1913-1921.
27. Scott-Browne, J.P., et al., *Evolutionarily Conserved Features Contribute to αβ T Cell Receptor Specificity.* Immunity, 2011. **35**(4): p. 526-535.
28. Ding, Y.H., et al., *Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids.* Immunity, 1998. **8**(4): p. 403-11.
29. Ding, Y.H., et al., *Four A6-TCR/peptide/HLA-A2 structures that generate very different T cell signals are nearly identical.* Immunity, 1999. **11**(1): p. 45-56.
30. Feng, D., et al., *Structural evidence for a germline-encoded T cell receptor-major histocompatibility complex interaction 'codon'.* Nat Immunol, 2007. **8**(9): p. 975-83.
31. Mazza, C., et al., *How much can a T-cell antigen receptor adapt to structurally distinct antigenic peptides?* Embo Journal, 2007. **26**(7): p. 1972-1983.
32. Deng, L., et al., *Structural insights into the editing of germ-line-encoded interactions between T-cell receptor and MHC class II by Vα CDR3.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(37): p. 14960-14965.
33. Yin, Y., Y. Li, and R.A. Mariuzza, *Structural basis for self-recognition by autoimmune T-cell receptors.* Immunol Rev, 2012. **250**(1): p. 32-48.
34. Holland, S.J., et al., *The T-cell receptor is not hardwired to engage MHC ligands.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(45): p. E3111-E3118.
35. Artyomov, M.N., et al., *CD4 and CD8 binding to MHC molecules primarily acts to enhance Lck delivery.* Proc Natl Acad Sci U S A, 2010. **107**(39): p. 16916-21.
36. Dong, D., et al., *Structural basis of assembly of the human T cell receptor-CD3 complex.* Nature, 2019. **573**(7775): p. 546-552.
37. Van Laethem, F., et al., *Deletion of CD4 and CD8 coreceptors permits generation of αβT cells that recognize antigens independently of the MHC.* Immunity, 2007. **27**(5): p. 735-750.
38. Burnet, F.M., *A modification of Jerne's theory of antibody production using the concept of clonal selection.* CA Cancer J Clin, 1976. **26**(2): p. 119-21.
39. Nossal, G.J., *Antibody production by single cells.* Br J Exp Pathol, 1958. **39**(5): p. 544-51.
40. Hozumi, N. and S. Tonegawa, *Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions.* Proc Natl Acad Sci U S A, 1976. **73**(10): p. 3628-32.
41. Glusman, G., et al., *Comparative genomics of the human and mouse T cell receptor loci.* Immunity, 2001. **15**(3): p. 337-49.
42. Parra, Z.E., et al., *Comparative genomic analysis and evolution of the T cell receptor loci in the opossum Monodelphis domestica.* BMC Genomics, 2008. **9**: p. 111.
43. Parra, Z.E., et al., *A unique T cell receptor discovered in marsupials.* Proc Natl Acad Sci U S A, 2007. **104**(23): p. 9776-81.
44. Connelley, T., et al., *Genomic analysis reveals extensive gene duplication within the bovine TRB locus.* Bmc Genomics, 2009. **10**.
45. Oettinger, M.A., et al., *Rag-1 and Rag-2, Adjacent Genes That Synergistically Activate V(D)J Recombination.* Science, 1990. **248**(4962): p. 1517-1523.
46. Mombaerts, P., et al., *Rag-1-Deficient Mice Have No Mature Lymphocytes-B and Lymphocytes-T.* Cell, 1992. **68**(5): p. 869-877.
47. Shinkai, Y., et al., *Rag-2-Deficient Mice Lack Mature Lymphocytes Owing to Inability to Initiate V(D)J Rearrangement.* Cell, 1992. **68**(5): p. 855-867.
48. Barreto, V., R. Marques, and J. Demengeot, *Early death and severe lymphopenia caused by ubiquitous expression of the Rag1 and Rag2 genes in mice.* Eur J Immunol, 2001. **31**(12): p. 3763-72.
49. Agrawal, A., Q.M. Eastman, and D.G. Schatz, *Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system.* Nature, 1998. **394**(6695): p. 744-51.
50. Max, E.E., J.G. Seidman, and P. Leder, *Sequences of five potential recombination sites encoded close to an immunoglobulin kappa constant region gene.* Proc Natl Acad Sci U S A, 1979. **76**(7): p. 3450-4.

51. van Gent, D.C., D.A. Ramsden, and M. Gellert, *The RAG1 and RAG2 proteins establish the 12/23 rule in V(D)J recombination.* Cell, 1996. **85**(1): p. 107-13.
52. Bassing, C.H., et al., *Recombination signal sequences restrict chromosomal V(D)J recombination beyond the 12/23 rule.* Nature, 2000. **405**(6786): p. 583-6.
53. Livak, F., et al., *Genetic modulation of T cell receptor gene segment usage during somatic recombination.* J Exp Med, 2000. **192**(8): p. 1191-6.
54. Taccioli, G.E., et al., *Impairment of V(D)J recombination in double-strand break repair mutants.* Science, 1993. **260**(5105): p. 207-10.
55. Komori, T., et al., *Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes.* Science, 1993. **261**(5125): p. 1171-5.
56. Ma, Y., et al., *Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination.* Cell, 2002. **108**(6): p. 781-94.
57. Rosati, E., et al., *Overview of methodologies for T-cell receptor repertoire analysis.* BMC Biotechnol, 2017. **17**(1): p. 61.
58. Weischenfeldt, J., et al., *NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements.* Genes & Development, 2008. **22**(10): p. 1381-1396.
59. Manfras, B.J., D. Terjung, and B.O. Boehm, *Non-productive human TCR β chain genes represent V-D-J diversity before selection upon function:: Insight into biased usage of TCRBD and TCRBJ genes and diversity of CDR3 region length.* Human Immunology, 1999. **60**(11): p. 1090-1100.
60. Smirnova, A.O., et al., *The use of non-functional clonotypes as a natural calibrator for quantitative bias correction in adaptive immune receptor repertoire profiling.* Elife, 2023. **12**.
61. Oestreich, K.J., et al., *Regulation of TCRβ gene assembly by a promoter/enhancer holocomplex.* Immunity, 2006. **24**(4): p. 381-391.
62. Stanhope-Baker, P., et al., *Cell type-specific chromatin structure determines the targeting of V(D)J recombinase activity in vitro.* Cell, 1996. **85**(6): p. 887-97.
63. Ji, Y., et al., *The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci.* Cell, 2010. **141**(3): p. 419-31.
64. Mathieu, N., et al., *Chromatin remodeling by the T cell receptor (TCR)-β gene enhancer during early T cell development:: Implications for the control of TCR-β locus recombination.* Journal of Experimental Medicine, 2000. **192**(5): p. 625-636.
65. Alt, F.W., et al., *Ordered Rearrangement of Immunoglobulin Heavy-Chain Variable Region Segments.* Embo Journal, 1984. **3**(6): p. 1209-1219.
66. Jhunjhunwala, S., et al., *Chromatin architecture and the generation of antigen receptor diversity.* Cell, 2009. **138**(3): p. 435-48.
67. Malissen, M., et al., *Regulation of TCR alpha and beta gene allelic exclusion during T-cell development.* Immunol Today, 1992. **13**(8): p. 315-22.
68. Thompson, S.D., J. Pelkonen, and J.L. Hurwitz, *First T cell receptor alpha gene rearrangements during T cell ontogeny skew to the 5' region of the J alpha locus.* J Immunol, 1990. **145**(7): p. 2347-52.
69. Hawwari, A., C. Bock, and M.S. Krangel, *Regulation of T cell receptor α gene assembly by a complex hierarchy of germline J*
promoters. Nature Immunology, 2005. **6**(5): p. 481-489.
70. Villey, I., et al., *Defect in rearrangement of the most 5' TCR-J alpha following targeted deletion of T early alpha (TEA): Implications for TCR alpha locus accessibility.* Immunity, 1996. **5**(4): p. 331-342.
71. Petrie, H.T., et al., *Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes.* J Exp Med, 1993. **178**(2): p. 615-22.
72. Wang, F., C.Y. Huang, and O. Kanagawa, *Rapid deletion of rearranged T cell antigen receptor (TCR) Valpha-Jalpha segment by secondary rearrangement in the thymus: role of continuous rearrangement of TCR alpha chain gene and positive selection in the T cell repertoire formation.* Proc Natl Acad Sci U S A, 1998. **95**(20): p. 11834-9.
73. Padovan, E., et al., *Expression of two T cell receptor alpha chains: dual receptor T cells.* Science, 1993. **262**(5132): p. 422-4.

74. Sarukhan, A., et al., *Allelic inclusion of T cell receptor alpha genes poses an autoimmune hazard due to low-level expression of autospecific receptors.* Immunity, 1998. **8**(5): p. 563-70.
75. Zal, T., et al., *Expression of a second receptor rescues self-specific T cells from thymic deletion and allows activation of autoreactive effector function.* Proc Natl Acad Sci U S A, 1996. **93**(17): p. 9102-7.
76. Yang, L., et al., *TCRα reporter mice reveal contribution of dual TCRα expression to Tcell repertoire and function.* Proceedings of the National Academy of Sciences of the United States of America, 2020. **117**(51): p. 32574-32583.
77. Pancer, Z. and M.D. Cooper, *The evolution of adaptive immunity.* Annu Rev Immunol, 2006. **24**: p. 497-518.
78. Finstad, J. and R.A. Good, *The Evolution of the Immune Response. 3. Immunologic Responses in the Lamprey.* J Exp Med, 1964. **120**(6): p. 1151-68.
79. Herrin, B.R. and M.D. Cooper, *Alternative Adaptive Immunity in Jawless Vertebrates.* Journal of Immunology, 2010. **185**(3): p. 1367-1374.
80. Ohno, S., *Evolution by gene duplication*. 1970, London, New York,: Allen & Unwin; Springer-Verlag. xv, 160 p.
81. Kapitonov, V.V. and J. Jurka, *RAG1 core and V(D)J recombination signal sequences were derived from transposons.* Plos Biology, 2005. **3**(6): p. 998-1011.
82. Fugmann, S.D., et al., *An ancient evolutionary origin of the Rag1/2 gene locus.* Proc Natl Acad Sci U S A, 2006. **103**(10): p. 3728-33.
83. Sakano, H., et al., *Sequences at the Somatic Recombination Sites of Immunoglobulin Light-Chain Genes.* Nature, 1979. **280**(5720): p. 288-294.
84. Koonin, E.V. and M. Krupovic, *Evolution of adaptive immunity from transposable elements combined with innate immune systems.* Nature Reviews Genetics, 2015. **16**(3): p. 184-192.
85. Fugmann, S.D., *The origins of the Rag genes--from transposition to V(D)J recombination.* Semin Immunol, 2010. **22**(1): p. 10-6.
86. Liu, C., et al., *Structural insights into the evolution of the RAG recombinase.* Nature Reviews Immunology, 2022. **22**(6): p. 353-370.
87. Olivieri, D.N., S. Gambon-Cerda, and F. Gambon-Deza, *Evolution of V genes from the TRV loci of mammals.* Immunogenetics, 2015. **67**(7): p. 371-84.
88. Nei, M., X. Gu, and T. Sitnikova, *Evolution by the birth-and-death process in multigene families of the vertebrate immune system.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(15): p. 7799-7806.
89. Clark, S.P., et al., *Comparison of human and mouse T-cell receptor variable gene segment subfamilies.* Immunogenetics, 1995. **42**(6): p. 531-40.
90. Fink, P.J. and M.J. Bevan, *H-2 antigens of the thymus determine lymphocyte specificity.* J Exp Med, 1978. **148**(3): p. 766-75.
91. Kyewski, B. and L. Klein, *A central role for central tolerance.* Annu Rev Immunol, 2006. **24**: p. 571-606.
92. Thome, J.J.C., et al., *Long-term maintenance of human naive T cells through in situ homeostasis in lymphoid tissue sites.* Science Immunology, 2016. **1**(6).
93. Douek, D.C., et al., *Changes in thymic function with age and during the treatment of HIV infection.* Nature, 1998. **396**(6712): p. 690-695.
94. Palmer, S., et al., *Thymic involution and rising disease incidence with age.* Proc Natl Acad Sci U S A, 2018. **115**(8): p. 1883-1888.
95. Palmer, D.B., *The effect of age on thymic function.* Frontiers in Immunology, 2013. **4**.
96. den Braber, I., et al., *Maintenance of Peripheral Naive T Cells Is Sustained by Thymus Output in Mice but Not Humans.* Immunity, 2012. **36**(2): p. 288-297.
97. Hazenberg, M.D., et al., *T cell receptor excision circles as markers for recent thymic emigrants: basic aspects, technical approach, and guidelines for interpretation.* Journal of Molecular Medicine-Jmm, 2001. **79**(11): p. 631-640.
98. Sempowski, G.D., et al., *T cell receptor excision circle assessment of thymopoiesis in aging mice.* Mol Immunol, 2002. **38**(11): p. 841-8.
99. Radtke, F., et al., *Deficient T cell fate specification in mice with an induced inactivation of Notch1.* Immunity, 1999. **10**(5): p. 547-58.
100. Swain, S.L., *T cell subsets and the recognition of MHC class.* Immunol Rev, 1983. **74**: p. 129-42.

References

101. Robey, E. and B.J. Fowlkes, *Selective events in T cell development.* Annu Rev Immunol, 1994. **12**: p. 675-705.
102. Shinkai, Y., et al., *Restoration of T cell development in RAG-2-deficient mice by functional TCR transgenes.* Science, 1993. **259**(5096): p. 822-5.
103. Levelt, C.N. and K. Eichmann, *Receptors and signals in early thymic selection.* Immunity, 1995. **3**(6): p. 667-72.
104. Saint-Ruf, C., et al., *Analysis and expression of a cloned pre-T cell receptor gene.* Science, 1994. **266**(5188): p. 1208-12.
105. Murata, S., et al., *Regulation of CD8+ T cell development by thymus-specific proteasomes.* Science, 2007. **316**(5829): p. 1349-53.
106. Nakagawa, T., et al., *Cathepsin L: critical role in Ii degradation and CD4 T cell selection in the thymus.* Science, 1998. **280**(5362): p. 450-3.
107. Gommeaux, J., et al., *Thymus-specific serine protease regulates positive selection of a subset of CD4 thymocytes.* European Journal of Immunology, 2009. **39**(4): p. 956-964.
108. Xing, Y., S.C. Jameson, and K.A. Hogquist, *Thymoproteasome subunit-beta5T generates peptide-MHC complexes specialized for positive selection.* Proc Natl Acad Sci U S A, 2013. **110**(17): p. 6979-84.
109. Ma, A., et al., *Bclx regulates the survival of double-positive thymocytes.* Proc Natl Acad Sci U S A, 1995. **92**(11): p. 4763-7.
110. Ueno, T., et al., *CCR7 signals are essential for cortex-medulla migration of developing thymocytes.* J Exp Med, 2004. **200**(4): p. 493-505.
111. Stritesky, G.L., et al., *Murine thymic selection quantified using a unique method to capture deleted T cells.* Proc Natl Acad Sci U S A, 2013. **110**(12): p. 4679-84.
112. Derbinski, J., et al., *Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels.* Journal of Experimental Medicine, 2005. **202**(1): p. 33-45.
113. Nagamine, K., et al., *Positional cloning of the APECED gene.* Nature Genetics, 1997. **17**(4): p. 393-398.
114. Gallegos, A.M. and M.J. Bevan, *Central tolerance to tissue-specific antigens mediated by direct and indirect antigen presentation.* Journal of Experimental Medicine, 2004. **200**(8): p. 1039-1049.
115. Voboril, M., et al., *Toll-like receptor signaling in thymic epithelium controls monocyte-derived dendritic cell recruitment and Treg generation.* Nature Communications, 2020. **11**(1).
116. Koble, C. and B. Kyewski, *The thymic medulla: a unique microenvironment for intercellular self-antigen transfer.* Journal of Experimental Medicine, 2009. **206**(7): p. 1505-1513.
117. Bouillet, P., et al., *BH3-only Bcl-2 family member Bim is required for apoptosis of autoreactive thymocytes.* Nature, 2002. **415**(6874): p. 922-926.
118. Xing, Y., et al., *Late stages of T cell maturation in the thymus involve NF-kappaB and tonic type I interferon signaling.* Nat Immunol, 2016. **17**(5): p. 565-73.
119. Davis, M.M. and P.J. Bjorkman, *T-cell antigen receptor genes and T-cell recognition.* Nature, 1988. **334**(6181): p. 395-402.
120. Thierry Mora, A.W., *Quantifying lymphocyte receptor diversity.* Systems Immunology, (ed. J. Das, C. Jayaprakash), 2019.
121. Murugan, A., et al., *Statistical inference of the generation probability of T-cell receptors from sequence repertoires.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(40): p. 16161-16166.
122. Quigley, M.F., et al., *Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(45): p. 19414-19419.
123. Doherty, P.C., J.M. Riberdy, and G.T. Belz, *Quantitative analysis of the CD8*
T-cell response to readily eliminated and persistent viruses. Philosophical Transactions of the Royal Society B-Biological Sciences, 2000. **355**(1400): p. 1093-1101.
124. Arstila, T.P., et al., *A direct estimate of the human αβ T cell receptor diversity.* Science, 1999. **286**(5441): p. 958-961.
125. Casrouge, A., et al., *Size estimate of the alpha beta TCR repertoire of naive mouse splenocytes.* J Immunol, 2000. **164**(11): p. 5782-7.
126. Langman, R.E. and M. Cohn, *The E-T (elephant-tadpole) paradox necessitates the concept of a unit of B-cell function: the protection.* Mol Immunol, 1987. **24**(7): p. 675-97.

127. Zarnitsyna, V.I., et al., *Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire.* Frontiers in Immunology, 2013. **4**.
128. Faint, J.M., et al., *Quantitative flow cytometry for the analysis of T cell receptor Vβ chain expression.* Journal of Immunological Methods, 1999. **225**(1-2): p. 53-60.
129. Pannetier, C., et al., *The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments.* Proc Natl Acad Sci U S A, 1993. **90**(9): p. 4319-23.
130. Redmond, D., A. Poran, and O. Elemento, *Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq.* Genome Med, 2016. **8**(1): p. 80.
131. Shugay, M., et al., *Towards error-free profiling of immune repertoires.* Nature Methods, 2014. **11**(6): p. 653-+.
132. Egorov, E.S., et al., *Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers.* Journal of Immunology, 2015. **194**(12): p. 6155-6163.
133. Dziubianau, M., et al., *TCR Repertoire Analysis by Next Generation Sequencing Allows Complex Differential Diagnosis of T Cell-Related Pathology.* American Journal of Transplantation, 2013. **13**(11): p. 2842-2854.
134. Robins, H.S., et al., *Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells.* Blood, 2009. **114**(19): p. 4099-4107.
135. Howie, B., et al., *High-throughput pairing of T cell receptor alpha and beta sequences.* Sci Transl Med, 2015. **7**(301): p. 301ra131.
136. Wang, C.L., et al., *High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(4): p. 1518-1523.
137. Carlson, C.S., et al., *Using synthetic templates to design an unbiased multiplex PCR assay.* Nature Communications, 2013. **4**.
138. Peng, Q., et al., *Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes.* Bmc Genomics, 2015. **16**.
139. *Rapid amplification of 5' complementary DNA ends (5' RACE).* Nat Methods, 2005. **2**(8): p. 629-30.
140. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nat Protoc, 2014. **9**(1): p. 171-81.
141. Freeman, J.D., et al., *Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing.* Genome Research, 2009. **19**(10): p. 1817-1824.
142. Das, D. and M.M. Georgiadis, *The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus.* Structure, 2004. **12**(5): p. 819-29.
143. Wulf, M.G., et al., *Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other.* J Biol Chem, 2019. **294**(48): p. 18220-18231.
144. Barennes, P., et al., *Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases.* Nat Biotechnol, 2021. **39**(2): p. 236-245.
145. Pieter Meysmana, et al., *Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report.* ImmunoInformatics, 2023. **9**.
146. Pai, J.A. and A.T. Satpathy, *High-throughput and single-cell T cell receptor sequencing technologies.* Nat Methods, 2021. **18**(8): p. 881-892.
147. Han, A., et al., *Linking T-cell receptor sequence to functional phenotype at the single-cell level.* Nature Biotechnology, 2014. **32**(7): p. 684-+.
148. Williams, R., et al., *Amplification of complex gene libraries by emulsion PCR.* Nature Methods, 2006. **3**(7): p. 545-550.
149. Turchaninova, M.A., et al., *Pairing of T-cell receptor chains via emulsion PCR.* Eur J Immunol, 2013. **43**(9): p. 2507-15.
150. Cusanovich, D.A., et al., *Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing.* Science, 2015. **348**(6237): p. 910-4.
151. Cusanovich, D.A., et al., *A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.* Cell, 2018. **174**(5): p. 1309-1324 e18.
152. Rosenberg, A.B., et al., *Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding.* Science, 2018. **360**(6385): p. 176-182.

References

153. Ma, S., et al., *Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin.* Cell, 2020. **183**(4): p. 1103-1116 e20.
154. Brown, S.D., L.A. Raeburn, and R.A. Holt, *Profiling tissue-resident T cell repertoires by RNA sequencing.* Genome Med, 2015. **7**: p. 125.
155. Jivanjee, T., et al., *Enriching and Characterizing T Cell Repertoires from 3' Barcoded Single-Cell Whole Transcriptome Amplification Products.* Methods Mol Biol, 2022. **2574**: p. 159-182.
156. Li, B., et al., *Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA-seq data.* Nat Genet, 2017. **49**(4): p. 482-483.
157. Tu, A.A., et al., *TCR sequencing paired with massively parallel 3' RNA-seq reveals clonotypic T cell signatures.* Nat Immunol, 2019. **20**(12): p. 1692-1699.
158. Frankish, A., et al., *GENCODE: reference annotation for the human and mouse genomes in 2023.* Nucleic Acids Res, 2023. **51**(D1): p. D942-D949.
159. Nurk, S., et al., *The complete sequence of a human genome.* Science, 2022. **376**(6588): p. 44-53.
160. Warren, R.L., et al., *Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes.* Genome Research, 2011. **21**(5): p. 790-797.
161. Fisher, R.A., A.S. Corbet, and C.B. Williams, *The relation between the number of species and the number of individuals in a random sample of an animal population.* Journal of Animal Ecology, 1943. **12**: p. 42-58.
162. Whittaker, R.H., *Vegetation of the Siskiyou Mountains, Oregon and California.* Ecological Monographs, 1960. **30**(3): p. 280-338.
163. Chiffelle, J., et al., *T-cell repertoire analysis and metrics of diversity and clonality.* Current Opinion in Biotechnology, 2020. **65**: p. 284-295.
164. Laydon, D.J., C.R.M. Bangham, and B. Asquith, *Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2015. **370**(1675).
165. Shannon, C.E., *A Mathematical Theory of Communication.* Bell System Technical Journal, 1948. **27**(3): p. 379-423.
166. Jaccard, P., *THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.* New Phytologist, 1912. **11**(2): p. 37-50.
167. Glanville, J., et al., *Identifying specificity groups in the T cell receptor repertoire.* Nature, 2017. **547**(7661): p. 94-98.
168. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
169. Dash, P., et al., *Quantifiable predictive features define epitope-specific T cell receptor repertoires.* Nature, 2017. **547**(7661): p. 89-93.
170. Steinegger, M. and J. Soding, *Clustering huge protein sequence sets in linear time.* Nat Commun, 2018. **9**(1): p. 2542.
171. Huang, H., et al., *Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening.* Nat Biotechnol, 2020. **38**(10): p. 1194-1202.
172. Arden, B., *Conserved motifs in T-cell receptor CDR1 and CDR2: implications for ligand and CD8 co-receptor binding.* Curr Opin Immunol, 1998. **10**(1): p. 74-81.
173. Bosc, N. and M.P. Lefranc, *The mouse (Mus musculus) T cell receptor alpha (TRA) and delta (TRD) variable genes.* Dev Comp Immunol, 2003. **27**(6-7): p. 465-97.
174. Bolotin, D.A., et al., *MiXCR: software for comprehensive adaptive immunity profiling.* Nat Methods, 2015. **12**(5): p. 380-1.
175. Kuchenbecker, L., et al., *IMSEQ-a fast and error aware approach to immunogenetic sequence analysis.* Bioinformatics, 2015. **31**(18): p. 2963-2971.
176. Li, S., et al., *IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling.* Nat Commun, 2013. **4**: p. 2333.
177. Liao, Y., G.K. Smyth, and W. Shi, *The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote.* Nucleic Acids Research, 2013. **41**(10).
178. Sidhom, J.W., et al., *DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires (vol 12, 1605, 2021).* Nature Communications, 2021. **12**(1).

179. Weber, A., J. Born, and M.R. Martínez, *TITAN: T-cell receptor specificity prediction with bimodal attention networks.* Bioinformatics, 2021. **37**: p. I237-I244.
180. Lu, T.S., et al., *Deep learning-based prediction of the T cell receptor-antigen binding specificity.* Nature Machine Intelligence, 2021. **3**(10): p. 864-+.
181. Luu, A.M., et al., *Predicting TCR-Epitope Binding Specificity Using Deep Metric Learning and Multimodal Learning.* Genes, 2021. **12**(4).
182. Hudson, D., et al., *Can we predict T cell specificity with digital biology and machine learning?* Nature Reviews Immunology, 2023. **23**(8): p. 511-521.
183. Springer, M.S., et al., *Placental mammal diversification and the Cretaceous-Tertiary boundary.* Proc Natl Acad Sci U S A, 2003. **100**(3): p. 1056-61.
184. (BfR), B.f.R., *Verwendung von Versuchstieren im Berichtsjahr 2022*. 2023.
185. Geliebter, J. and S.G. Nathenson, *Recombination and the Concerted Evolution of the Murine Mhc.* Trends in Genetics, 1987. **3**(4): p. 107-112.
186. Silberman, D., et al., *Class II major histocompatibility complex mutant mice to study the germ-line bias of T-cell antigen receptors.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(38): p. E5608-E5617.
187. Nobuhara, H., et al., *Polymorphism of T-Cell Receptor Genes among Laboratory and Wild Mice - Diverse Origins of Laboratory Mice.* Immunogenetics, 1989. **30**(6): p. 405-413.
188. Pullen, A.M., et al., *Surprisingly uneven distribution of the T cell receptor V beta repertoire in wild mice.* J Exp Med, 1990. **171**(1): p. 49-62.
189. Migalska, M., A. Sebastian, and J. Radwan, *Profiling of the TCRβ repertoire in non-model species using high-throughput sequencing.* Scientific Reports, 2018. **8**.
190. Abolins, S.R., et al., *Measures of immune function of wild mice,.* Molecular Ecology, 2011. **20**(5): p. 881-892.
191. Lochmiller, R.L., M.R. Vestey, and S.T. Mcmurry, *Primary Immune-Responses of Selected Small Mammal Species to Heterologous Erythrocytes.* Comparative Biochemistry and Physiology a-Physiology, 1991. **100**(1): p. 139-143.
192. Masopust, D., C.P. Sivula, and S.C. Jameson, *Of Mice, Dirty Mice, and Men: Using Mice To Understand Human Immunology.* Journal of Immunology, 2017. **199**(2): p. 383-388.
193. Mestas, J. and C.C.W. Hughes, *Of mice and not men: Differences between mouse and human immunology.* Journal of Immunology, 2004. **172**(5): p. 2731-2738.
194. Beura, L.K., et al., *Normalizing the environment recapitulates adult human immune traits in laboratory mice.* Nature, 2016. **532**(7600): p. 512-6.
195. Graham, A.L., *Naturalizing mouse models for immunology.* Nat Immunol, 2021. **22**(2): p. 111-117.
196. Gregorova, S. and J. Forejt, *PWD/Ph and PWK/Ph inbred mouse strains of Mus m. musculus subspecies--a valuable resource of phenotypic variations and genomic polymorphisms.* Folia Biol (Praha), 2000. **46**(1): p. 31-41.
197. Rajabi-Maham, H., et al., *The south-eastern house mouse Mus musculus castaneus (Rodentia: Muridae) is a polytypic subspecies.* Biological Journal of the Linnean Society, 2012. **107**(2): p. 295-306.
198. Dejager, L., C. Libert, and X. Montagutelli, *Thirty years of Mus spretus: a promising future.* Trends Genet, 2009. **25**(5): p. 234-41.
199. Frazer, K.A., et al., *A sequence-based variation map of 8.27 million SNPs in inbred mouse strains.* Nature, 2007. **448**(7157): p. 1050-3.
200. Liu, K.J., et al., *Interspecific introgressive origin of genomic diversity in the house mouse.* Proc Natl Acad Sci U S A, 2015. **112**(1): p. 196-201.
201. Staubach, F., et al., *Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (Mus musculus).* PLoS Genet, 2012. **8**(8): p. e1002891.
202. Dureje, L., et al., *The mouse hybrid zone in Central Europe: from morphology to molecules.* Folia Zoologica, 2012. **61**(3-4): p. 308-318.
203. Jing, M., et al., *Phylogeography of Chinese house mice (Mus musculus musculus/castaneus): distribution, routes of colonization and geographic regions of hybridization.* Mol Ecol, 2014. **23**(17): p. 4387-405.
204. Keane, T.M., et al., *Mouse genomic variation and its effect on phenotypes and gene regulation.* Nature, 2011. **477**(7364): p. 289-94.

205. Guenet, J.L. and F. Bonhomme, *Wild mice: an ever-increasing contribution to a popular mammalian model.* Trends Genet, 2003. **19**(1): p. 24-31.
206. Beck, J.A., et al., *Genealogies of mouse inbred strains.* Nature Genetics, 2000. **24**(1): p. 23-+.
207. Villa-Morales, M., J. Santos, and J. Fernandez-Piqueras, *Functional Fas (Cd95/Apo-1) promoter polymorphisms in inbred mouse strains exhibiting different susceptibility to gamma-radiation-induced thymic lymphoma.* Oncogene, 2006. **25**(14): p. 2022-9.
208. Staelens, J., et al., *Hyporesponsiveness of SPRET/Ei mice to lethal shock induced by tumor necrosis factor and implications for a TNF-based antitumor therapy.* Proc Natl Acad Sci U S A, 2002. **99**(14): p. 9340-5.
209. Graham, J.B., et al., *Extensive Homeostatic T Cell Phenotypic Variation within the Collaborative Cross.* Cell Reports, 2017. **21**(8): p. 2313-2325.
210. Tang, F.C., et al., *mRNA-Seq whole-transcriptome analysis of a single cell.* Nature Methods, 2009. **6**(5): p. 377-U86.
211. Munsky, B., G. Neuert, and A. van Oudenaarden, *Using gene expression noise to understand gene regulation.* Science, 2012. **336**(6078): p. 183-7.
212. Raj, A. and A. van Oudenaarden, *Nature, nurture, or chance: stochastic gene expression and its consequences.* Cell, 2008. **135**(2): p. 216-26.
213. Ramaswamy, S., et al., *Multiclass cancer diagnosis using tumor gene expression signatures.* Proc Natl Acad Sci U S A, 2001. **98**(26): p. 15149-54.
214. Chan, Y.F., et al., *Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Enhancer.* Science, 2010. **327**(5963): p. 302-305.
215. Carroll, S.B., *Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.* Cell, 2008. **134**(1): p. 25-36.
216. Wittkopp, P.J. and G. Kalay, *Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence.* Nat Rev Genet, 2011. **13**(1): p. 59-69.
217. Kornberg, R.D. and Y. Lorch, *Chromatin structure and transcription.* Annu Rev Cell Biol, 1992. **8**: p. 563-87.
218. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.
219. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nat Methods, 2013. **10**(12): p. 1213-8.
220. Baysoy, A., et al., *The technological landscape and applications of single-cell multi-omics.* Nature Reviews Molecular Cell Biology, 2023. **24**(10): p. 695-713.
221. Ma, S., et al., *Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin.* Cell, 2020. **183**(4): p. 1103-+.
222. Rothenberg, E.V., J.E. Moore, and M.A. Yui, *Launching the T-cell-lineage developmental programme.* Nature Reviews Immunology, 2008. **8**(1): p. 9-21.
223. Jaeger, E.E., R.E. Bontrop, and J.S. Lanchbury, *Structure, diversity, and evolution of the T-cell receptor VB gene repertoire in primates.* Immunogenetics, 1994. **40**(3): p. 184-91.
224. Zerrahn, J., W. Held, and D.H. Raulet, *The MHC reactivity of the T cell repertoire prior to positive and negative selection.* Cell, 1997. **88**(5): p. 627-36.
225. Soares, P., et al., *Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock.* American Journal of Human Genetics, 2009. **84**(6): p. 740-759.
226. Marrack, P., et al., *Evolutionarily conserved amino acids that control TCR-MHC interaction.* Annual Review of Immunology, 2008. **26**: p. 171-203.
227. Nowak, M.A., K. Tarczyhornoch, and J.M. Austyn, *The Optimal Number of Major Histocompatibility Complex-Molecules in an Individual.* Proceedings of the National Academy of Sciences of the United States of America, 1992. **89**(22): p. 10896-10899.
228. Woelfing, B., et al., *Does intra-individual major histocompatibility complex diversity keep a golden mean?* Philosophical Transactions of the Royal Society B-Biological Sciences, 2009. **364**(1513): p. 117-128.
229. Migalska, M., A. Sebastian, and J. Radwan, *Major histocompatibility complex class I diversity limits the repertoire of T cell receptors.* Proc Natl Acad Sci U S A, 2019. **116**(11): p. 5021-5026.
230. Flajnik, M.F., *The immune system of ectothermic vertebrates.* Veterinary Immunology and Immunopathology, 1996. **54**(1-4): p. 145-150.

231. Pierini, F. and T.L. Lenz, *Divergent Allele Advantage at Human MHC Genes: Signatures of Past and Ongoing Selection.* Molecular Biology and Evolution, 2018. **35**(9): p. 2145-2158.
232. Penn, D.J., K. Damjanovich, and W.K. Potts, *MHC heterozygosity confers a selective advantage against multiple-strain infections.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(17): p. 11260-11264.
233. Forsberg, L.A., et al., *Influence of genetic dissimilarity in the reproductive success and mate choice of brown trout - females fishing for optimal MHC dissimilarity.* Journal of Evolutionary Biology, 2007. **20**(5): p. 1859-1869.
234. Jacob, S., et al., *Paternally inherited HLA alleles are associated with women's choice of male odor.* Nat Genet, 2002. **30**(2): p. 175-9.
235. Reusch, T.B.H., et al., *Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism.* Nature, 2001. **414**(6861): p. 300-302.
236. Messaoudi, I., et al., *Direct link between polymorphism, T cell avidity, and diversity in immune defense.* Science, 2002. **298**(5599): p. 1797-1800.
237. Borghans, J.A.M., A.J. Noest, and R.J. De Boer, *Thymic selection does not limit the individual MHC diversity.* European Journal of Immunology, 2003. **33**(12): p. 3353-3358.
238. Selin, L.K., S.R. Nahill, and R.M. Welsh, *Cross-Reactivities in Memory Cytotoxic T-Lymphocyte Recognition of Heterologous Viruses.* Journal of Experimental Medicine, 1994. **179**(6): p. 1933-1943.
239. Wucherpfennig, K.W. and J.L. Strominger, *Molecular Mimicry in T-Cell-Mediated Autoimmunity - Viral Peptides Activate Human T-Cell Clones Specific for Myelin Basic-Protein.* Cell, 1995. **80**(5): p. 695-705.
240. West, J.D., et al., *Development of Interspecific Hybrids of Mus.* Journal of Embryology and Experimental Morphology, 1977. **41**(Oct): p. 233-243.
241. Potts, W.K. and P.R. Slev, *Pathogen-based models favoring MHC genetic diversity.* Immunol Rev, 1995. **143**: p. 181-97.
242. Woodland, D.L., B.L. Kotzin, and E. Palmer, *Functional Consequences of a T-Cell Receptor D-Beta-2 and J-Beta-2 Gene Segment Deletion.* Journal of Immunology, 1990. **144**(1): p. 379-385.
243. Nanda, N.K., R. Apple, and E. Sercarz, *Limitations in plasticity of the T-cell receptor repertoire.* Proc Natl Acad Sci U S A, 1991. **88**(21): p. 9503-7.
244. Li, H., et al., *Determinants of public T cell responses.* Cell Res, 2012. **22**(1): p. 33-42.
245. Gras, S., et al., *Allelic polymorphism in the T cell receptor and its impact on immune responses.* J Exp Med, 2010. **207**(7): p. 1555-67.
246. Cabaniols, J.P., et al., *Most α/β T cell receptor diversity is due to terminal deoxynucleotidyl transferase.* Journal of Experimental Medicine, 2001. **194**(9): p. 1385-1390.
247. Russell, M.L., et al., *Combining genotypes and T cell receptor distributions to infer genetic loci determining V(D)J recombination probabilities.* Elife, 2022. **11**.
248. Fazilleau, N., et al., *Vα and V public repertoires are highly conserved in terminal deoxynucleotidyl transferase-deficient mice.* Journal of Immunology, 2005. **174**(1): p. 345-355.
249. Srivastava, S.K. and H.S. Robins, *Palindromic Nucleotide Analysis in Human T Cell Receptor Rearrangements.* Plos One, 2012. **7**(12).
250. Fuschiotti, P., et al., *Analysis of the TCR α-chain rearrangement profile in human T lymphocytes.* Molecular Immunology, 2007. **44**(13): p. 3380-3388.
251. Correia-Neves, M., et al., *The shaping of the T cell repertoire.* Immunity, 2001. **14**(1): p. 21-32.
252. Posnett, D.N., et al., *Level of Human Tcrbv3s1 (V-Beta-3) Expression Correlates with Allelic Polymorphism in the Spacer Region of the Recombination Signal Sequence.* Journal of Experimental Medicine, 1994. **179**(5): p. 1707-1711.
253. Mark, M., et al., *Viral infection reveals hidden sharing of TCR CDR3 sequences between individuals.* Frontiers in Immunology, 2023. **14**.
254. Argaet, V.P., et al., *Dominant Selection of an Invariant T-Cell Antigen Receptor in Response to Persistent Infection by Epstein-Barr-Virus.* Journal of Experimental Medicine, 1994. **180**(6): p. 2335-2340.
255. Day, E.K., et al., *Rapid CD8 T cell repertoire focusing and selection of high-affinity clones into memory following primary infection with a persistent human virus:: Human Cytomegalovirus.* Journal of Immunology, 2007. **179**(5): p. 3203-3213.

256. Trautmann, L., et al., *Selection of T cell clones expressing high-affinity public TCRs within human cytomegalovirus-specific CD8 T cell responses.* Journal of Immunology, 2005. **175**(9): p. 6123-6132.
257. Davison, A.J., *Evolution of the herpesviruses.* Veterinary Microbiology, 2002. **86**(1-2): p. 69-88.
258. Madi, A., et al., *T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity.* Genome Research, 2014. **24**(10): p. 1603-1612.
259. Merbl, Y., et al., *Newborn humans manifest autoantibodies to defined self molecules detected by antigen microarray informatics.* Journal of Clinical Investigation, 2007. **117**(3): p. 712-718.
260. Kidman, J., et al., *Characteristics of TCR Repertoire Associated With Successful Immune Checkpoint Therapy Responses.* Frontiers in Immunology, 2020. **11**.
261. Hopkins, A.C., et al., *T cell receptor repertoire features associated with survival in immunotherapy-treated pancreatic ductal adenocarcinoma.* Jci Insight, 2018. **3**(13).
262. Page, D.B., et al., *Deep Sequencing of T-cell Receptor DNA as a Biomarker of Clonally Expanded TILs in Breast Cancer after Immunotherapy (vol 4, pg 835, 2016).* Cancer Immunology Research, 2017. **5**(3): p. 269-269.
263. van Heijst, J.W.J., et al., *Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation.* Nature Medicine, 2013. **19**(3): p. 372-377.
264. Simoni, Y., et al., *Bystander CD8 T cells are abundant and phenotypically distinct in human tumour infiltrates.* Nature, 2018. **557**(7706): p. 575-+.
265. Li, B., et al., *Landscape of tumor-infiltrating T-cell repertoire of human cancers.* Cancer Research, 2016. **76**.
266. Springer, I., N. Tickotsky, and Y. Louzoun, *Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction.* Frontiers in Immunology, 2021. **12**.
267. Kamga, L., et al., *CDR3α drives selection of the immunodominant Epstein Barr virus (EBV) BRLF1-specific CD8 T cell receptor repertoire in primary infection.* Plos Pathogens, 2019. **15**(11).
268. Carter, J.A., et al., *Single T Cell Sequencing Demonstrates the Functional Role of alphabeta TCR Pairing in Cell Lineage and Antigen Specificity.* Front Immunol, 2019. **10**: p. 1516.